THE AMERICAN ECONOMIC REVIEW

VOLUME LXVII

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

Editorial Office: Brown University, Providence, Rhode Island

CONTENTS

Monetary Theory for Open Economies: The State of the Art

Recent Controversies in Monetary Theory

Welfare Economics

Equilibrium in Markets Where Price Exceeds Cost

Application of Microsimulation Methodology

British Capital in the Late Nineteenth Century: Sources in Britain and Movement in the Empire

Impact of Recent Developments in Public Finance Theory on Public Policy Decisions

Ethics in Government

Environmental Problems

Exhaustible Resources

PROCEEDINGS

The purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. Persons of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

THIS volume contains the papers and proceedings of the eighty-ninth annual meeting of the American Economic Association. "Proceedings" concern the business activities of the Association—the annual membership meeting, the twice yearly meetings of the Executive Committee, and reports of various Association officers and committees—and, as with the "Notes" section in each issue of the *American Economic Review*, are published in the hope that members will be informed of, and will participate in, the Association's affairs.

The "Papers," which constitute the greater part of this volume, are roughly equivalent to two regular issues of the *American Economic Review*, but are published under quite different procedures.

About a year in advance, the Association's President-elect (in 1976 Lawrence R. Klein, in 1977 Jacob Marschak) sets in motion the process which culminates in the presentation at the annual meetings and publication of the papers. After consultation and comments, both volunteered and solicited, from a wide range of persons, the President-elect decides on the topics of sessions at which papers will be presented, sets limits on the length of papers and invites persons to organize these sessions. Each session organizer in turn invites, again after consultation, several persons (usually two or three in number) to give papers on the topic of the session in question, and asks others to give comments on the papers. The papers, or rough drafts, are supposed to be sent to the editors of the *Papers and Proceedings* a month and a half before the meetings; we check them and where appropriate send the authors comments and suggestions. The papers and comments are read at the Association's meetings; we ask the authors to send us the final versions of their papers within three days. After giving the final versions further scrutiny, we send them to the printer to be published. We do not generally publish discussants' comments unless there is some special reason to do so.

In 1976 there was one significant departure from the procedures of previous years. Besides the usual "invited paper sessions" (sixteen in number in 1976 plus the Richard T. Ely Lecture), there was a substantial number of "contributed paper sessions." The Program Committee selected the authors on the basis of abstracts submitted in response to a notice in the *American Economic Review*, and the papers themselves were considered for publication in the *Papers and Proceedings* by the Program Committee. Also considered for publication were papers given at sessions sponsored jointly by the American Economic Association and another member of the Allied Social Science Associations (called "joint paper sessions"). Of the nearly one hundred papers considered for publication, there was space for less than one in four. Inevitably a substantial number of papers of publishable quality had to be omitted. Where two or more of the successful papers in the competition were on similar subjects, we have grouped them under a general title whether they were in

fact given at the same session or not. In a few instances, we have added a contributed paper to a group of invited papers as if it had been given at the same session. At the end of the *Papers* section of this issue, we have grouped several papers under the general heading "Selected Contributed Papers."

Thus, the rules under which these papers are published are quite different from those concerning articles appearing in regular issues of the *Review*. Most of the papers published here are invited, not contributed; except in unusual circumstances they must be less than 4,000 (and sometimes 3,000 or even 2,000) words in length; each author is asked to avoid highly technical discussion and instead to contribute a paper which will be intelligible to economists not specializing in the subject it discusses. While we do edit most papers to improve content and style and to satisfy space constraints, we do not subject invited papers to any refereeing process, and publication of any invited paper which we receive prior to the final printer's deadline and which satisfies (or can be cut to satisfy) space requirements is virtually guaranteed.[1]

These policies are advantageous in a number of respects. The papers can be published without the long delays inherent in publication by regular journals. They are relatively short papers which concern a wide variety of subjects, and can be understood by nonspecialists. Authors get a chance to report briefly on research just completed, to discuss topical subjects in an informal way, and to summarize longer forthcoming publications; readers get a chance to browse among a large number of articles which are outside their major areas of interest but which are not as specialized or ponderous as those sometimes found in regular journals. And while the papers are not scrutinized in the way regular journal articles are, they are (or are at least intended to be!) nontechnical, and thus do not need the thorough refereeing usual for articles submitted to journals.

RENDIGS FELS
C. ELTON HINSHAW

[1]We would refuse to publish such a paper only if we felt that it was utterly without merit, no paper has yet been rejected on these grounds. However, the Executive Committee has established another ground for rejection. If a paper cannot be cut to meet space requirements, we ask the author either to authorize us to submit it to the managing editor of the *Review* to be considered for publication (subject to the usual refereeing process) in a regular issue of that journal, or to withdraw the paper and submit it elsewhere

RICHARD T. ELY LECTURE

Two Centuries of Economic Growth: Reflections on U.S. Experience

By SIMON KUZNETS*

Over most of the two past centuries, the country's growth was a movement from the small, largely agricultural, economy of thirteen divided colonies on the Atlantic shores, to a unified, industrialized, advanced economy of continental dimensions. The combination of a high rate of growth of population, peopling an expanding territory, with a rise in per capita product and productivity associated with a stream of technological innovations and rapid shifts in the structure of production, was of uniquely large impact in the United States; and while sharing much in common with the growth experience of other, currently developed, countries, displayed several distinctive features.

Reflecting on this process, one may raise four questions. First, how was the high rate of population growth attained? Second, how was the impressive rate of growth in per capita product sustained over most of the period, at least back to the early 19th century? Third, how, in the face of divisive sectional interests and differential impacts of rapid economic growth, was unity preserved and economic inequality affected? Fourth, how does one evaluate the drastic shifts that occurred since World War I in the international framework within which this country lived and grew? Such broad questions, and their implications, can be treated only briefly and incompletely; but, at least, they may help to organize the discussion.

I. The Growth of Population

In the mid-1770's, the population of the thirteen original colonies was 2.5 million. At that time, the population of Great Britain was 9 million; of France over 24 million; of Europe, excluding Russia, 128 million (all within the 1914 boundaries). By 1910, the population of the United States was 91.6 million (excluding, for comparability, the minor group of nonwhites other than Negroes)—over twice that of either Great Britain or France (each with about 40 million at that date). By mid-1975, the population of the United States was close to 214 million. The multiplication factor over the two centuries was over 85 for the *U.S.* population; for Europe, including or excluding European Russia, it was about 4; for the more rapidly growing among the European countries, not much more than 7. Nor was the contrast less striking in comparison with the population growth of Japan.

The contrast is, of course, the result of a long period of cumulation of the excess of annual or decennial rates of increase in the United States. Over the two centuries, the average rate of population growth *per year* was about 2¼ percent in the United States, and 0.9 percent in Great Britain. Taken over a decade or two, the cumulative difference would not be large; taken over two centuries, it cumulated to a contrast in multiples between over 85 and barely over 6.

A high population growth rate can be due to a high birth rate, or to a low death rate, or to a high net in-migration rate. In comparison with other developed countries, specifically those in Europe, the death rates, crude or refined, were not that much lower in this country as to contribute significantly to the much greater capacity of *U.S.* population to grow. The sources of the difference lay largely in the birth and in-migration rates. The birth rate in the United States in the early 19th century was estimated at

*Professor of Economics Emeritus, Harvard University.

close to 50 per thousand—high even by current standards in the less developed countries. While it dropped rapidly in the early decades of the last century and moved further downwards to the low levels of today, it tended to remain distinctly higher than in the older developed countries—as was also the case in the other, young, overseas countries like Canada, Australia, and New Zealand.

The other major source of growth was immigration. For the country's black population, immigration, or rather importation, was of significant proportions between 1770 and 1810, but negligible thereafter. For the far larger white population, immigration contributed little between the Revolution and the mid-1830's, but was an important factor over the eight decades prior to World War I and for a few years after the war. The magnitude of this immigration (on which we shall concentrate henceforth) can be illustrated by references to the millions of immigrants who came in and stayed. But one must also take account of the offspring of these immigrants in succeeding generations, a net addition if we assume that the birth rates of the original, nonimmigrant population would have declined just as much, if not more, without immigration. The assumption is reasonable, since we find substantial declines in native white birth rates in periods (before 1840) and later also in regions (South and rural non-South), that were relatively little affected by white immigration. A calculation, made by a Census expert (W. S. Rossiter) using the native white birth rates prevailing in the past, estimated the contribution of the original white stock (i.e., the one in this country at the time of the Revolution) to the 1920 population of the United States at 47.3 million out of a total white population of 94.8 million—the rest being immigrants and their descendants. This result, that half of the population was to be credited to immigrants and their descendants, would be roughly valid for total population, including the Negroes; and would also hold true of the more recent dates after 1920.

The persistence, over two centuries, of birth rates higher than those in the older developed countries of Europe and Japan, and the prevalence over some eight to nine decades of net immigration that contributed so much to population increase, may reasonably be associated with the "newness" of the United States. The "newness" meant the presence, and an awareness of the presence, of vast resources in unsettled land of a geographically expanding nation; a population that had detached itself from the European economic and institutional constraints on early marriage and on prolific child bearing; a willingness to welcome and encourage immigration, if, after a while, only within limits as to the cultural areas of origin. These features of the United States as a new country, the first among the overseas offshoots of Europe to achieve political sovereignty, are well known and hardly need documentation, even were it feasible here. But we might note aspects that seem relevant to understanding some distinctive characteristics of this country's growth.

The first comment relates to the long period over which obviously large reserves of land and resources remained and could continue to exercise effects on birth and immigration rates. The land area of the country within the continental United States (i.e. except Hawaii and Alaska), as it became fixed in the 20th century, amounted, at the censuses of 1790 and 1800, to 0.82 million square miles, grew by the census of 1810 to 1.7 million miles (reflecting the Louisiana Purchase) and to 2.94 million, the present size, after the census of 1840. A more telling series relates to the area of *settled* land, land with a population density of at least 2 persons per square mile. From 1790 to 1890, this area, originally 0.24 million square miles, grew at an average of over 50 percent every twenty years, including a growth of 53 percent from 1870 to 1890. Even when the limit was reached at 1.9 million in 1890, the closing of the frontier did not mean the absence of room for expansion. A similar story of a long process of settlement of a wide, and for a long while expanding, territorial base is told by the dates of admission of new

states into the Union: the last in the continental United States were admitted as late as 1907 and 1912—over a century and a quarter after the founding of the Republic. That native birth rates began declining so early is no reflection of a pressure of limits, but indicates a sensitivity to increasing affluence and to greater density in the *older* settled parts of the country; or, conversely, to the costs of internal migration toward the still abundant land.

Second, the desire to people the continent, to use the country's sovereignty to extend its area and to add to its population without interference from a metropolitan, colonial monarchy, was explicitly indicated in the Declaration of Independence—with its reference to the attempt on the part of the British monarch "to prevent the population of these States; for that purpose obstructing the Laws of Naturalization of foreigners; refusing to pass others to encourage their migration hither...." And this declared willingness, subject later to some selective restrictions, to admit immigrants from a wide range of European countries and those in the Americas, persisted until shortly after World War I. Thus, for some three-quarters of the two-century span, the United States was a country of open immigration, the only one with so long a record and with a huge absorptive capacity combined with a high standard of living.

In considering the impact of immigration, one must keep in mind its selectivity. There was selectivity not only with respect to age and sex, which resulted in a high proportion of males in prime working ages; not only with respect to occupations and attachments within the country of origin, which made for high proportions of farmers and farm workers, common laborers and semiskilled artisans; not only with respect to individual characteristics that favored the more adventurous and adaptable among the younger groups within the labor force; not only with respect to timing, which meant that, with the exception of the 1842 Irish famine, the impelling occasion was the entry of the sending country into modern economic growth, with its dislocating effects on rural population and that part of the urban population that might have been adversely affected by industrialization. There was also selectivity of immigration flows in their location, geographical and socio-structural, in the United States, the country of destination.

To begin with, few immigrants settled in the South, a census region largely identical with the slave-owning states and extending from Maryland-Delaware in the North to Florida in the South, and from the Atlantic coast states in the East to Texas and Oklahoma in the West. Already by 1860, when the foreign born were close to 19 percent of all whites in the regions outside the South, the percentage in the South was below 6; an allowance for native born of foreign or mixed parentage would raise the proportion of the foreign white stock to about 28 percent of the whites in the non-South and only 8½ percent in the South. This selectivity persisted, so that in 1910, when the foreign white stock proportion (close to 50 percent of all whites) was at its peak in the non-South, that proportion in the South was below 10 percent. Second, the immigrant flow tended towards the urban rather than rural areas, and to the bigger cities within the urban. Thus, already in 1850, the first year of data on the foreign born, the proportion of foreign born whites (and of free colored, a small component) to the corresponding total in the non-South was 14 percent; but it was 39 percent in the large cities, and 11 percent in the other city areas. By 1920, the foreign white stock accounted for 66 percent of total white population in cities of 500,000 and over; about 44 percent in cities of 25 to 500 thousand; 35 percent in the smaller cities; and only 20 percent of the rural population. Third, while this concentration in the urban areas and under-representation in the rural meant that the foreign stock, both foreign born and the first generation of their descendants, were under-represented in agriculture, there was selectivity even in the urban occupations, at least among the foreign born. For obvious reasons they were concentrated in the laborers and operatives categories, and under-represented among the white

collar and professional pursuits.

Only a few tentative remarks can be made here on the impact of immigration on the country's economic growth. The direct economic effect was to add to the labor supply, particularly in the non-South, the addition representing human capital investment made in the country of origin; and to provide an inflow of labor to urban and related pursuits at costs presumably lower than those that would have been involved in attracting the native labor force to move from older pursuits in the countryside and smaller cities. The more indirect economic effects lay in making possible a more rapid growth to a larger scale, with whatever special economies and efficient production possibilities such more rapid growth and larger scale may have implied. The wider, less narrowly economic effects, lay in assisting to tilt the balance of power against the slave-holding South, white immigration representing an effective vote for the free labor, industrializing economy; in diversifying the cultural and historical sources of the American population, and affirming the role of the United States as the haven and long-term base of populations, dislocated, particularly in Europe, in the transition from the preindustrial to the industrial economy; and, in placing on the educational and cultural institutions of the country the task of assimilating the newcomers, and especially their children, into the community. In general, one may suggest that because of the availability of immigration, the *U.S.* economy and society were able to operate with a wider range of choice—at least in that part of the country to which the immigration flowed freely and in significant numbers.

The drastic reduction of immigration, combined with the continuing secular decline in the rate of natural increase, brought the period of an impressively high population growth to an end by the late 1920's. From 1790 to 1830, the rate was at 29.4 per thousand per year, with immigration negligible. From 1830 to 1885, the rate was 26.8 per thousand per year, and net immigration accounted for as much as 6.0 points, or almost a quarter. From 1885 to 1925 the rate was 18.1 per thousand per year, with immigration accounting for 4.8 points, or over a quarter. For 1925–1970, the rate was only 12.6 per thousand per year—still significantly higher than in most older developed countries—with immigration accounting for only 1.2 points, or less than a tenth.

The slowing down of population growth, and the drastic decline in the flow of economically oriented immigrant labor to the country, had a variety of consequences. Some of them are touched upon below. Here we can refer briefly to some effects of immigration restriction. Internally, it meant, once the worst of the depression of the 1930's was over, that the reduction in the inflow of immigrant labor opened up opportunities for more employment of native labor at similar skill levels, particularly of Negroes from the South. It is hardly an accident that while the proportion of all Negroes in the country residing in the South hovered at about 90 percent from 1770 to 1910, it began to decline with World War I, and by 1970 dropped to 53 percent, with substantial shares of Negro population appearing in the other regions, particularly the North. Conversely, the proportion of foreign born in the total white population of the non-South, at a peak of about 21 percent in 1910, dropped to below 6 percent in 1970. Likewise, the cessation of mass immigration, flowing in the past largely to the middle-Atlantic shores, must have affected differentials in population growth among the various regions of the country. And, of course, immigration restriction signalled the end of the United States as an open country, as a haven for economically displaced workers and population from Europe and elsewhere.

II. Growth in Per Capita and Total Product

The high growth rate of population in the United States was combined with a substantial growth rate in per capita product. In shifting from numbers of people to the magnitude of the output that they turned out, we face the complexities of the economic and social coverage of net product (or gross of capital consumption

only) and its valuation. One has to recognize that the magnitudes are affected by the price scales applied, the use of initial prices yielding higher growth rates than the use of terminal prices; that omission of some production in kind will impart an upward bias to growth rates; and so on. But we are concerned here with rough orders of magnitude, employing linked indexes of series in which quantities are weighted by changing price ratios.

In looking back to the 1770's, we find that the record to 1800 yields a rather uncertain result, with the period affected by revolution, war, the immediate following difficulties, and recovery. For 1800 to 1840 we have tentative estimates, which can be accepted as suggesting growth of about 1 percent per capita per year—a substantial rate of growth by the standards of the time. For the next forty-five years, from 1834–43 to 1879–88, the rate, based on totals inclusive of improvements in kind, and manufacturing value-added, in agriculture, was between 1.3 and 1.5 percent per year. It then rose over the next two periods—from 1880–89 to 1920–29, and from 1920–29 to 1970—to between 1.6 and 1.8 percent per year. If, to secure a simple cumulative result, we assume that there was no growth in per capita product between 1770 and 1800, and cumulate over the remaining years from 1800 to 1970, we find that per capita product rose by a factor of somewhat over 11½ (over the two centuries). Before comparing it with the record for other developed countries, we should note that with the growth in population by a factor of over 85 and in per capita product of over 11½, the scale of the economy, as reflected in total product, must have grown by a factor close to 1,000. It is this latter figure that recapitulates the movement of the United States from a small, largely agrarian economy, two hundred years ago, to the huge, industrialized economy of today.

In attempting comparison with other countries, we encounter difficulties in that the records for most other countries do not go back as far as for the United States; and, more important, that one should expect a higher growth rate in per capita product in a country that enters the phase of industrialization and modern economic growth later. We can compare the United States with Great Britain-United Kingdom, for the stretch back to 1800 or to later initial dates; and overall, the growth rate in per capita product in the United States is distinctly higher, by perhaps a quarter. In comparisons with France (back to 1840) and Germany (back to 1850), we find the rates for the three countries fairly similar. Higher growth rates are found in the Scandinavian countries, particularly Sweden, the comparisons beginning in the 1860's; in Japan, the comparison beginning in the 1800's; and in Italy, the comparison beginning in the 1890's. But we know that in Italy and Japan the earlier periods in the 19th century were marked by low growth in per capita product; and the same may have been true of the Scandinavian countries prior to the 1860's, although we have no relevant evidence at hand. Hence, if the comparison between the United States and these several countries with higher growth rates in per capita product in the more recent (if still long) periods, were extended back to, say, 1800, the differences would most likely disappear, or be reversed. The suggestion, of more general relevance, is that a later entry into modern economic growth, assuming that the growth is then sustained, is associated with higher rates of increase in per capita product once growth begins; and extension to longer periods in the comparisons for countries that have attained an adequate level of development, reduces differences associated with the *timing* of the start. This *making-up* characteristic of modern economic growth is found also in other sequences (e.g., in connection with the differential impact of a war, or of other interruptions in the "normal" course of growth).

A study for 1970 (by Irving Kravis and others), based on detailed analysis of comparative prices, yields a per capita product for the United States that, in terms of international prices, exceeds that of the United Kingdom by a ratio of 100 to 60; of France and Germany by a ratio of 100 to 75; and is about equal to that of

Sweden (with rough allowance made here for differences between exchange rate and international price conversions). Extrapolation of such ratios back by per capita growth rates in the United States and in other countries yields a relationship between the *initial* per capita products in the international prices of 1970, or in some hybrid set of prices if chain indexes of product adjusted by price indexes to different time bases were used. A direct comparison in the international prices of, say, 1800, or 1840, or 1870, might look different. But the calculation still permits a judgment that the initial levels of per capita product in the United States were comparatively high, even before industrialization proceeded far. Indeed, it is doubtful that per capita product in this country in the early 19th century was much lower than that in Great Britain, the leading industrial country of the world at the time (the shortfall could hardly have been more than a fifth, if that); and it was clearly above the initial per capita product of the other European countries, which entered the process of industrialization in the 1840's or later. Thus, the United States, in the early 19th century (and the late 18th) was an agricultural country, but productive and rich. One of the sources of its quantitative dominance in the economic world of later and more recent decades was that the high growth rates of its population were combined with substantial growth rates of per capita product sustained over a long period *and* applied to an initial per capita income of a level that was already high.

One should have expected substantial growth in per capita product in this country, its major source being that associated with modern economic growth—i.e., technological advance, connected in varying degrees of closeness with the advance of science and useful knowledge. After all, the American revolution came about the same time as the industrial revolution. Great Britain, the original mother country, was, through most of the 19th century, the leader in the industrial revolution; and the major technological breakthroughs connected with the textile and chemical industries, with the iron and steel industries, and the introduction of steam power, were easily accessible to and found prompt application in this country. Indeed, the United States, through most of the century, was noted for effective adaptation and modification of the advancing world technology to fit it better to the country's resource endowments; and then later, in the electric and internal combustion age, began to contribute more heavily to the initial inventions and innovations. In the still more recent period, beginning shortly after World War I—a period of some five decades marked by extraordinary advances in health, agriculture, the spread of internal combustion to air transport and of electricity to household services, the emergence and spread of the electronic and nuclear revolutions, and so on to space exploration—the United States played a far more active and leading role than it had in the technological revolutions of the century and a half that preceded World War I.

We are so used to sustained and substantial growth in per capita product that we tend to take it for granted—not realizing how exceptional growth of that magnitude is on the scale of human history; and how much it requires in the complicated process of invention, application, accumulation, and adjustment. If we find that, say, over a quarter of a century, per capita product rose by 50 percent, this means that usually with the same or smaller labor input per capita, the working population managed to produce that much more of final product—food, clothing, shelter etc., *and* whatever additional capital, material or human, was needed to produce it. Such a feat can be accomplished either because of a lucky gift of hitherto unused natural resources—hardly a sustainable source, except through advance of knowledge that creates resources out of hitherto useless components of nature; or because of greater learning, within the context of already available knowledge—again a quickly exhaustible source without creation of new knowledge that extends the limits within which learning can occur; or, and most importantly, because of new inventions, which, when applied, enlarge the productive

capacity of human labor. And, indeed, when one looks behind the rather unrevealing economic aggregates, one finds a stream of technological changes representing the applications of new inventions and new knowledge—and contributing, when applied, to further learning, discovery, and invention. A glance at a single sector in the United States, say that of internal transport, reveals a sequence of canals and turnpikes, steamboats on internal waterways and steam railroads, electric railroads, internal combustion engine transport and highways, air transport—all of this in successive major breakthroughs, and cycles of emergence, learning, expansion, and eventually obsolescence.

Technological innovations, which constitute the major permissive source of modern economic growth, carry constraints of their own, even in a country like the United States that also enjoyed extensive expansion and access to additional natural resources. The innovations require, for effective application, specific responses from the society desirous of utilizing them. And these, in turn, mean adjustments in economic and social institutions, differential impacts on various groups within a society, and effects on even purely economic relations, e.g., the amounts of capital investment that have to be generated to embody the technological innovation, relative to the net product that it will yield. Thus, the domestic capital formation proportions that we find in the United States in the 19th century—at over 20 percent gross or close to 15 percent net—were substantially higher than the 10 to 12 percent gross in Great Britain or the Scandinavian countries at the time; and may be viewed as responses to the capital demanding infrastructure of residential and related construction, railroads and other public utilities, in a continental country, with a rapidly growing population. One may also note in the reproducible capital stock at the end of the century the high proportion of capital in transport, communication, housing and related construction—the capital investment in manufacturing and agriculture becoming proportionately greater only later. And the completion of capital-demanding infrastructure in the 19th century, and the marked slowing down in the growth of population and labor in the 20th century, may perhaps explain the greater rate of growth of factor productivity in the recent decades—with a less capital-demanding technology.

But the effects of technological innovations were not only on capital formation and factor productivity. They were also on the organization of economic production or management units, in the pressure for the modern type of corporation; and they had a ramifying effect on industrial organization through the use of the discriminating power of monopoly. They affected conditions of work, with changes in labor force status, employment requirements, educational levels, and the active lifespan of the working population; and they affected conditions of life, through furthering urbanization and modifying patterns of consumption and other elements in the modes of living associated with rising economic standards. The various institutional adjustments, and shifts in conditions of work and life, required for effective channeling of the continuous stream of technological innovations, were neither easy, nor costless. The gap between the stock of knowledge and inventions as the necessary condition, and the institutional and social adjustments that would convert the former into a sufficient condition, is wide—as past history of the economically developed countries and the current history of the less developed amply show. That the United States achieved a sustained and fairly high rate of growth of per capita product over this long period is evidence of the country's capacity to modify its institutions and patterns of work and life, at rates sufficient to accommodate the technological potentials and in ways that preserved, except for the Civil War, a freely accepted social consensus.

The emphasis on the technological innovations, associated with a growing stock of knowledge, involves the implicit argument that conventional measures of factor productivity, even if expanded to include investment in

human capital, are incomplete. This is so at least at present, when our understanding of the processes by which new knowledge and new inventions originate is so meager, and so long as the economic calculus is of limited application to a resource the returns from which are so wide-flung in space and time, and the identifiable costs of which are in such disproportion to returns when observable. One should also add that the feedback effects of the application of new inventions in mass production on the facilitation of additional knowledge and invention have not been studied sufficiently to provide an adequate body of data. Do we really understand, in economic terms, the succession of various sources of industrial power, and can we explain, e.g., the timing of the emergence of the electronic revolution in communication? Questions such as these are pertinent to the analysis of *U.S.* growth even in the 19th century, when the United States was a follower country applying largely European discoveries and inventions. They become of critical significance in the recent decades when this country has attained sufficient leadership to become itself the major source of advance in new knowledge and invention.

III. Unity and Inequality

The political and social framework of a country sets the major conditions for economic growth, in formulating and monitoring rules of economic and social behavior; and changing them, when adjustments are required by new obstacles and opportunities brought by accumulated costs of the past, new knowledge, and new external circumstances. Since modern economic growth means a succession of differential impacts of innovations on different groups within a society, unified, effective decisions may be required to preserve consensus, minimize negative impacts and maximize the positive contributions of growth. Indeed, a major function of modern sovereign government is to help channel social and political adjustments to economic growth, to modify old and create new institutional patterns that would facilitate growth while limiting its inequitable effects. Given the variety of, and likely conflicts among, the group interests affected, an overriding sovereign power is required that would represent the interests and values of the community.

The problem of maintaining flexible and creative unity despite divisiveness produced by modern economic growth, was complicated in the case of the United States by several historical circumstances. To begin with, the nation was formed of thirteen colonies, which, by the time the new political entity began operating, had had well over a century of separate existence, and thus opportunity to develop different economic, political, and social characteristics. The distinction between the North and the South (more specifically the Northeast and the Southeast) was sharply marked, already in 1790—the year of the first census and within the country's first presidential term. In that year, of 1.97 million population in the North, only 3 percent were Negroes, and of these fewer than two-thirds were slaves; in the South, of a similar total population of 1.96 million, over 35 percent were Negroes and of these over 95 percent were slaves. One can also find data on the tonnage of trade of the various colonies in 1770, which clearly point to the dominance in the North of trade with the West Indies, and in the South of trade with Great Britain. The subsequent persistence of the original North-South cleavage, and its sharpening to a clash between incompatible bases of economic and social organization, led to a civil war almost a century after the American revolution. While the legal abolition of slavery marked, in one way, the end of this clash, the heritage persisted in the isolation of the South and the continued economic and social discrimination against the Negro—not to be effectively mitigated until the post-World War II decades.

Next, even setting aside the conflict with regard to slavery, a long period of political experimentation and innovation was required to weld the original, and increasing, number of states into an organization capable of formulat-

ing and enforcing unified decision, and, indeed, of establishing the common interests that these decisions were intended to serve. At least three novel elements were involved, setting the conditions in which the evolution of a unified country had to take place. First, there was the basic decision to launch a new nation by agreement among former colonies that declared an end to their old allegiance to a single, outside, authority. This, in itself, represented a revolutionary novelty, a major innovation; and like all major innovations, it needed prolonged experimentation and adjustment before it could attain a realistically optimum level. The period of such adjustment would have been long even without additional complications of rapid geographic expansion, and, after an early date, of intensive industrialization and major technological advance. But, second, this new nation, with only emergent unifying powers and only gradually widening bases for common action, was in the process of rapid westward expansion, with special sets of problems created by the movement of people to the frontiers and the addition of new state units to the older commonwealth. The emergence, and conditions of admission, of these new units were of differing consequence to the several older parts of the country; and while such geographical expansion provided a strong sense of unity to the country, the specific changes had to be made without too much damage to the consensus. Third, and most relevant to economic growth, there was the process of industrialization and structural transformation, a flow of novel changes requiring new institutional and legal patterns, and affecting differently the several groups in the population. There was, consequently, need for some single authority, acting for the country and capable of evolving—to monitor and select the necessary institutional and legal adjustments, and try to provide the proper channels for economic advance while mitigating its adverse effects.

The results of *U.S.* economic growth are clearly seen in the high rates of growth of population and of per capita product the process could also be viewed in a series of growth-setting decisions. These would begin with the commitment to political independence from outside, and political unity within; and would then involve the implementation of that independent and unifying power in a series of decisions—on the public domain, the treatment of debt, free labor and slavery, internal improvements, regulation of foreign trade, public education, and so on in a long list. It is not possible here, nor am I competent, to attempt such a list, in proper order and weight of decisions. One can only observe that the successive decades of the 19th and early 20th centuries witnessed a series of secular- or growth-decisions, the long-term implications of which were largely perceived by the different groups aware of their interests but also cognizant of some common goals; that, if one can judge by the changing political organization, the trend has been towards a continuing widening in popular participation, at least in the election of representatives charged with exploring, and arriving at, the decisions; and that, finally, at least prior to World War I, there seemed to have been a persistent thread in these growth decisions. The thread was provided by a desire to people the continental span of the country, and to exploit the large scale opportunities provided, on the one hand, by the stock of natural resources perceived as such in the light of current knowledge, and on the other, by the advance in modern technology which created new resources and widened markedly the range of productivity of human labor organized within an adequate social framework. Both extensive and intensive expansion was pursued, by a country open to immigration and unconcerned with external threats or, after the civil war, with dangers to internal unity.

Extensive expansion ceased at some time in the early 20th century, within a span of years extending from the closing of the frontier at the end of the 19th century, to the admission of the last state in the continental United States in 1912, to the effects of World War I of 1914-18, to the sharp restriction of immigration in the

mid-1920's. The period of five decades that followed was quite different; and even within it, there was a contrast between the first twenty-five years from the mid-1920's to the end of the 1940's—with a major depression and a world war, and the last quarter of a century. It is only during the latter subperiod that a variety of adjustments occurred, adjustments to the cessation of mass immigration with its differential impact on regions and on communities of different size, and to major shifts in world conditions.

In turning now to economic inequality, changes in which are a potent source of unity and disunity, I find it difficult to deal broadly with this wide and complex aspect of the country's economic growth. My interest is largely in inequality generated by economic growth, and the difficulty is in finding data and analysis that would cover both the growth-induced income disparities and the offsets through mobility—all of this with proper cognizance also of changes in family and household structure generated by modern demographic trends. But it may be useful to call attention to special elements in our historical experience, which differed between the long sweep to World War I and the more recent period since the late 1920's.

In the earlier period, the existence of slavery over the first century after independence, and of effective legal and social discrimination in the South in later decades, introduce elements that render conventional economic measures unrevealing and inadequate. Whatever shortfalls there were in the calculated economic returns to the people bound in slavery, or to those with sharply restricted rights, they were a limited part of the story; and the major part was hardly susceptible of a purely economic calculation. Here was a case of economic, legal, and social deprivation that persisted over three quarters of the total long-time span, and allowed only limited relief through mobility—all of this applying to a substantial group within the country's population. In greatly reduced form, the observation may apply even to the majority of the white population in the South, relative to the white population in the other regions. The former were not afforded opportunities as great as those for the white population elsewhere, since the slave population and its custom-bound successors prior to World War I failed to provide the domestic markets and thus the growing demand, that the local white population could satisfy and grow with. Nor were conditions in the South a good preparation for a would-be white migrant to regions outside the South, a fact that inhibited such migration.

Within the long period prior to the 1920's, in a subperiod beginning with the late 1830's, the income distribution among the population outside the South (almost all white) was complicated by the incidence of mass immigration. The latter, with the typically lower incomes of the foreign born, meant an addition of weight to the lower tail of the income distribution—even though, to the immigrant himself, the income, even in his earlier years in the country, may have meant a marked advance over what he was earning in his country of origin. And, most likely, this income-inequality-widening effect of the entry of immigrants varied over time with variations in the relative inflow and the widening contrast between the income levels prevailing in the United States and those at which employment openings could be filled by the newcomers. But the same factor also made for higher mobility up the income ladder—in that with the passage of time and accumulation of experience, the income of the foreign born would rise more rapidly than that of the native born; and in that, as the data indicate, the incomes of the next generation, native born of foreign parents, would show a rise over the incomes of their parents greater than between two successive generations of native born of native parents.

With substantial mobility of labor in and out of the country in the decades before World War I, there was only limited pressure for sustained government intervention to supplement income during depressions by unemployment compensation or public works; or to provide for old age pensions through governmental security plans.

And with the hoped-for mobility up the economic ladder, at least for the white population, under conditions of peace and rapid growth, there was no great pressure for governmental policy to reduce income inequalities, except through assurance of equality of opportunity. The impression I have is that the income distribution in the United States, in the decades before World War I and for some years thereafter—until the great depression of the 1930's—was little modified by government intervention.

To what extent the situation changed after the mid-1920's and particularly since the early 1950's, is a matter for exploration by scholars more familiar with the trends in this recent period. I can only offer conjectures. As already indicated, there has been in the recent period a movement of Negroes away from the South and to other regions; and there has been a marked advance in removing limitations and discrimination, particularly after World War II. This should have led to a reduction of economic differentials, and, most important, to a weakening of restrictions on opportunities and on mobility. The marked reduction in the volume of net immigration and the shift in its composition away from dominance by labor of lower skills, should have reduced its contribution to the low tail of the income distribution among the white population in the non-South. At the same time, it should have reduced mobility over time within the income distribution. But there may well have been offsetting changes elsewhere.

Most impressive was the marked trend toward greater government intervention, to provide some offsets to the incidence of income deficiency occasioned by unemployment, illness, breakdown within the family, and old age insecurity; and to extend equality of opportunity through enforcement of the rights of hitherto restricted minorities. The trend, emerging first during the depth of the depression of the 1930's, in response to critical levels of unemployment and economic deprivation, expanded much further after World War II. It was due only partly to the stabilization of the *U.S.* population and labor force, following the reduction of immigration; and it was due partly to the slowly shifting views on the peace-type goals of economic and social life. But it was also due to the realization that with the incidence and dangers of wars affecting the country and threatening its population, the burdens imposed by discrimination, and by the purely competitive pressures of the unregulated private market, should not be tolerated. There was, apparently, a line of connection between changes in the international framework within which this country had to operate after World War I and the policies of the government (later involved in the massive programs connected with defense) bearing on equality of opportunities and on income distribution.

IV. Recent Changes in the International Framework

By the international framework within which a country lives and grows I mean the structure of the rest of the world, with which the given country engages either in peaceful exchange of goods, men, capital, and ideas; or in active, or potential, conflict involving the use of force. At a given time, this structure of the rest of the world would differ from country to country, depending on its size, location, economic and social characteristics, and the like; and it would change over time for a given country, as the latter and the rest of the world change, and as the means of contact among them also change.

One omission, among several, in our selective discussion so far is the neglect of the salient and changing aspects of the international framework within which this country has been operating since the early days of its political independence. This omission cannot be repaired here: doing so would require coverage of the peaceful flows of trade, migration, and capital; of the conditions of tension and conflict in the rest of the world, and between some of it and this country; and of the changing technology of international relations. Yet, because of its obvious major impact on the structure of economic growth of this country in recent decades, one

should note briefly the marked change that occurred in the political and conflict aspects of the international framework as it may be perceived for this country.

World War I, coming after almost a century of relative peace (punctuated only by local wars), and followed within two decades by World War II, signified the beginning of a new period for the United States, as it did for many other nations. After withdrawal from European stresses and conflicts since the early 19th century, this country participated in both world conflicts; and modified its policies to suit the new conditions of growing world disarray. The very occurrence of "world" wars, i.e., ones characterized by prolonged and costly participation by a high proportion of the major developed countries of the world (together with some less developed partners), meant that, by the early 20th century, the number of such large industrialized countries had grown sufficiently large to have generated numerous points of conflict. It also suggests that, despite the obvious mutual advantages of growing volumes of peaceful trade and capital flows, there were sufficiently large elements of international competition and friction in modern economic growth under the auspices of increasingly nationalistic sovereign states, to make the occurrence of a war a high probability.

Several consequences of such major wars may be noted. First, and most direct, they accentuated the advance of war technology —which, however, in developed countries, is an integral part of the country's technological complex. Thus, the advance of technology since the late 18th century increased the capacity and productivity of long-distance transport and communication at least as much, if not more, than it did that of the production of commodities and other services. Modern technology bridged space gaps within and among countries that barred flows of goods and men for centuries; and it resulted, by the mid-20th century, in a world in which no part of mankind was really isolated from others (except, in some countries, by government fiat). But such revolutionary improvements in transport were just as important for delivery of war materiel and armies as they were for peace-type transport; and, indeed, the advance in the capacity of delivering war "goods" at long distance was clearly greater. Likewise, the increased technological power of mankind, i.e., the greater power to modify natural processes to satisfy human purposes, was perhaps as great, if not greater, when these purposes had to do with destruction in time of war than with construction for peaceful ends. Thus, the enormous advance in transport and communication resulted in economic and political interdependence among nations that was quite recent and new in the long history of human societies; and came after millenia of almost isolated existence, during which distinctive historical heritage was accumulated by different societies, little affected by, and indeed often unaware of, the rest of the world. But the removal of isolation meant also the removal of protection. For the United States, as for many other countries, protective (as well as inhibiting) distance from other powers shrank rapidly, particularly after World War II.

Second, participation in the prolonged and major conflicts meant, for the developed countries and their less economically developed partners, a strain that led often to political breakdowns and the emergence of new and deviant forms of political and social organization. In the less developed countries, like Russia and China, the heretofore gradually-growing modern elements were weakened by World Wars I and II, sufficiently to give way to Communism. Among the developed countries of Europe, the first World War led to the dissolution of a multinational monarchy like Austria-Hungary, and the emergence of fascism in Italy, Germany, and some of the other European states—another case of the use of a hierarchically organized dictatorial party to force the growth of economic and political power of the country by ideologically claimed control over the population. Since these were new approaches, representing violent breaks with the past, explicit hostility to the past, and to other

nations still associated with it and representing competing forms, became a long-term policy at times taking particularly virulent forms. These outbreaks of deviant and self-proclaimed revolutionary regimes, emerging as engines of accelerated political and economic growth, introduced into the world, particularly after the 1920's, elements of cleavage and divisiveness that were absent, or only latent, before World War I.

Finally, one should note that the world wars came as a result of the culmination of antecedent and competitive expansion by the economically developed countries towards colonization of much of the rest of the world. A consequence of World War I was to demonstrate that the advantages of such colonization to the developed countries were limited. And this demonstration was greatly reinforced by the realization that the tutelage of the colonies by the metropolitan countries was self-terminating if there was to be sharing of modern values—a sharing inevitable in continued contact. The shift was finally completed in the course of World War II when distant colonies were lost so easily; and when it became evident that, with the advance of modern technology, the advantages of presumably secure natural resources in the colonies were limited, while the rights of the native inhabitants of the colonies to be the masters of their own political and hence, presumably, also economic, destinies, were paramount. The result was a remarkable spread of national sovereignty extending to large numbers of hitherto colonial areas, to some after World War I, but to others at a far greater rate after World War II.

It may seem paradoxical that precisely at the time when technological progress broke down the isolation in the world and made for increased economic interdependence, divisive boundaries of national sovereign statehood spread so widely; and that there occurred a striking decentralization of political power among a mushrooming number of new and small jurisdictions. But perhaps this is not paradox at all. If the world has become so much tighter, and countries are exposed to both benefits and dangers from so many possible outside sources, a national society that shares a strong feeling of community of kind, might desire to have the freedom of sovereign decision to be exercised in crucial choices. And this would be all the more so, when the government is in the hands of a monolithic minority party that might want to have the power and trappings of sovereignty to protect itself internally, and to isolate the country from external influences viewed as temptation or corrupting knowledge.

As the comments above suggest, the world wars were only a reflection of the underlying causes that brought about the major shifts in the international framework since the 1920's, and particularly rapidly since the early 1950's. They reflect the enormous technological contribution in the developed countries, which was accessible to, and adopted by much of the rest of the world, but in a selective way; and they also reflect the strains and stresses that economic growth was creating in both developed and developing countries and that led to nationalistic and aggressive policies—with whatever ideological claims were evolved to justify the latter. Even without overt wars, the combination of advance in technological power, for good and bad, with its differential spread to, and impact on, countries at different stages of development, and with the shrinking of distance in the world, would have resulted in much greater international tension than existed in the earlier periods of greater distance and isolation.

Whatever the causes, and the comments above provide only tentative suggestions, the changes in the international framework in the recent decades—the increased divisiveness, more intensified ideologically-powered hostility, and the greater danger of war-induced devastation—involve heavy costs to this country, as well as to many others. These costs should be noted not only in terms of large military budgets, and the absorption of a larger proportion of high level scientific and technological manpower in war-related work. There are also the costs of distortion of channels of cooperation and communication in an ideologically di-

vided world; and the costs involved in the greater complexity within the country's economic and social organization, which must provide the means for viable policy decisions—both on the domestic use of the increased technological power for equitable economic and social advance, and on the problems of relations with the rest of the world that may be so explosive.

The growth problems of a developed country can be viewed within the context of a combination of technological and economic power, present and prospective; of a variety of accepted goals, and hence of responsibilities; and of the dangers of unforeseen (some unforeseeable) errors and of unavoided (some unavoidable) failures. One may characterize this combination for the United States, recently and currently, as that of enormous power, wide responsibilities, and substantial dangers. The very size of the country's population and economic product, and particularly the large reservoir of its scientific and technologically creative human resources, give it enormous power, currently and in prospect. The responsibilities are wide because the country's decisions—on the directions of basic and applied research, on policy with respect to agriculture and agricultural stockpiling, on nuclear and other energy, on weapon production and sales, on multinational corporations and so on in a long list—have a marked impact not only on its own population but also on much of the rest of the world. The dangers of error and failure are formidable because the power of advanced technology makes errors potentially that much more costly; because so much of the rest of the world needs assistance in its attempts to bridge the gap between attainments and minimum aspirations; and because the destructive potentialities of modern technology are so much greater, particularly in a divided world.

Within the past two centuries, and associated with modern economic growth, there must have been many such combinations of increasingly great technological and economic power, with diverse goals, and the greater dangers associated with errors and failures. Yet, even with lagging adjustments and costly failures, the results, at least in terms of material returns, showed a fairly marked upward trend. Even in this recent twenty-five year period of greater strain and danger, the growth in peace-type product per capita in the United States was still at a high rate; and in the rest of the world, developed and less developed (but excepting the few countries and periods marked by internal conflicts and political breakdown), material returns have grown, per capita, at a rate higher than that ever observed in the past. And one should note that current problems, still unresolved, always loom larger than those of the past—which have been resolved sufficiently for us to have survived and flourished and for us to be able to view them more dispassionately.

But long-term projections into ranges well beyond those covered by the observed past are subject to wide errors; and the variables and parameters under discussion (and many more should be cited) are too diverse and too crude to permit adequate analysis, certainly within the limits of my competence. The purpose of the brief comment was to emphasize the association between growth of technological and economic power (stemming in large part from new knowledge) occurring under the aegis of the nationalist sovereign state, and the probability of errors of innovation (based, by definition, on incomplete knowledge) and of international strain and conflict.

AMERICAN ECONOMIC GROWTH: IMPORTED OR INDIGENOUS?

What Difference Did the Beginning Make?

By J. R. T. HUGHES*

A subject as complex as this is best begun as simply as possible. Would we be as we are with other beginnings? Obviously not. Those countries started as European colonial enterprises that had different beginnings are now Canada, Mexico, Brazil, Cuba, etc. We are as we are largely because of the materials from which our society was initially formed. All were transmuted with time, but despite that the original ingredients would remain in the mixture and influence the long-term results. We are still identifiably similar to our colonial ancestors in our institutional structure and behavior regarding economic life. The colonists, in turn, had transplanted their own laws and practices from England. Intellectual and institutional continuity is thus a reality. It is also interesting and sometimes surprising.

I. The Colonial and English Background

Because of *Dartmouth College v. Woodward* (1819), one of the most celebrated cases in constitutional law, the English origin of the power of contract in the United States is a well-known legacy. Without English notions of contract everything from the earliest Royal patents to the lowliest contracts of labor indenture would have been different. But there was something else colonial and English in that famous court proceeding, and that was judicial review itself. *Marbury v. Madison* (1803) established the principle of judicial review at the federal level with the Supreme Court overturning an action of the federal government. Where did this idea come from? It was, curiously, something we got from the English, although it was not English practice. The charters of the colonies provided that legislative acts of colonial assemblies must not be repugnant to the laws of England. So for eight decades, from 1696 to 1776, legislative acts of the colonies were reviewed by the Board of Trade lawyers and some 5 percent of them were thrown out. At the Constitutional Convention in 1787 this practice was discussed and it was agreed that the laws of the states should be subject to federal review. It was the genius of John Marshall to establish just how that would be done. One need think only of the *Schechter* case overturning the National Industrial Recovery Act to appreciate the powerful economic consequences of this legacy.

Contract and judicial review by themselves would justify an essay such as this. But they are only the tip of the iceberg. Actually the entire common law of England made the passage to colonial America; the elements of economic society, land, labor, trade controls, business practice, bailment, the rules governing all were planted here. The accepted formula was that the law of any colony was the entire common law plus the English statute laws up to the time of colonization, together with those later statutes specifically extended by Parliament to the colonies. An important law like the *Statute of Frauds* of Charles II would be adopted by colonial assemblies outright, and after the Revolution would be repassed verbatim by the new state legislatures. The common law of England was claimed by the Continental Congress in 1774, and in new state constitutions after the Revolution began. Common law proceedings

*Professor of Economics, Northwestern University. The major themes discussed here are the subject of my book, *Social Control in the Colonial Economy*, and may there be pursued at length. In the pages that follow references are given for direct quotations only.

became primarily the concern of state courts, but there were some common law decisions by the Supreme Court of the United States, one of which contained the following warning against excessive originality. In *Robinson v. Campbell* (1818) the state courts were instructed:

> ... the remedies in the courts of the United States are to be, at common law or equity, not according to the practice of the state courts, but according to the principles of common law and equity, as distinguished and defined in that country from which we derive our knowledge of those principles

The United States in 1776 was about to begin the long journey of Revolution and independence, and even though the English were being thrown out, a selection of their laws was being taken along on the journey. As Chief Justice Morrison Waite put it in *Munn v. Illinois* (1877): "When the people of the United Colonies separated from Great Britain, they changed the form, but not the substance, of their government." The laws of England were the laws of the colonies, and subsequent American law and practice would grow from that intellectual root Since these rules included laws regulating economic life, I will turn to those specifically, and discuss their relevance. We will treat property rights in real estate, the labor contract, and crucial parts of general business practice. Out of these, ultimately, the behavioral rules of the American economy would emerge.

II. Land

In a country whose productive resources were destined to be privately owned the tenure (rights) of ownership in real property would be crucial. The way Americans own their real property came from a past as ancient as Saxon England. From the Virginia patent of 1606 onwards the crown prohibited all English tenures except that of free and common socage from those colonies which were destined to become the United States For long-run development this meant that: 1) a nation of family farms would develop; 2) exploitation of minerals and timber would be privately determined until the mid-20th century; 3) with common law in force the same would later be true of oil and natural gas, and water rights would develop in a way favoring private development interests. What would not occur would be large-scale ownership of idle lands, the latifundia of Spanish America. Moreover, the restrictions of the English land rules after Edward I would prohibit establishment of subinfeudation except in Maryland and Pennsylvania where, in any case, such did not flourish. The incidents of socage were fixed and certain, so they could be discounted and a fixed land price could be established by bargain. Socage inheritance was direct to the heir, or else by will, and the land could be freely aliened. Since the property right in socage land vanished if the incidents of the tenure were not met, the reserved rents, later state and local property taxes, would have to be paid, thus funding into the far future local government taxing authorities without question. Land was thus turned into a commodity. The necessity to meet the incidents of socage was understood by Adam Smith who foresaw the family farm economy to come as early as 1776.

> But in all the English colonies the tenure of the lands, which are all held by free socage, facilitates alienation, and the grantee of any extensive tract of land, generally finds it for his interest to alienate, as fast as he can, the greater part of it, reserving only a small quit rent. [p 539]

It is amusing to consider how many American economists and economic historians have read past those lines without understanding their meaning So far as I know, no modern economic history of the United States bothers to mention "free socage." Its establishment here was a fortunate externality (not without its problems) worth teaching those who have an historical interest in the *U.S.* economy. Socage was a crucial determinant of the country's future course of development. The tenure was never changed. We call it fee simple now, since it shared the crucial characteristics of rights to waste, free alienation and direct inheritance with the English fee simple of the ancient military tenures. Primogeniture and double portion

disappeared, and after the Ordinance of 1787 inheritance by equal portions in equal degrees of consanguinity spread. Long leases disappeared and entailments generally came to be limited to a single life only.

The place of the American Indian's right to the continent was also determined by the English and colonial regimes, and it changed little afterward, certainly little in favor of the Indians, until they finally were left with reservations, individual private ownership, and such residual group rights as are still under litigation.

III. The Labor Contract

The legacy of the labor contract was less happy than that of land ownership, but no less enduring in many respects. As was true of Indian land rights, the right of Black people to ownership of their labor power was originally extremely limited. But such was true of the majority of whites as well. Slavery meant a history of race trouble that lasts until now; the legacy of white indentured servitude left the laborer "below the salt" in American society despite federal establishment of organized labor in 1935.

One does not need to dwell now upon American Negro slavery; it has been celebrated extensively by American economists in recent years. It was established by the colonists, propagated by the English and until 1772 was allowed in England by law. Chief Justice Sir John Holt had ruled early in the 18th century that "negroes are merchandise and within the Navigation Acts," (McCrady, p. 385) but in 1772, in the Somerset case, Lord Mansfield argued that slavery was not compatible with the English constitution, leaving us with the remote possibility that had the English not got the American Revolution from us, they might have been presented with the Civil War instead.

As Abbot Smith said of labor relations in colonial America: "Colonial society was not democratic, and certainly not equalitarian; it was dominated by men who had money enough to make others work for them" (p. 7). In fact, simple compulsion lay at the root of the matter. Slaves were compelled to labor for life; white servants transported under bonds of indenture (apart from the Puritan migration, these were estimated to be from half to two-thirds of all white immigrants) were compelled to labor for fixed terms of years. Convicts were transported along with war prisoners as indentured servants. Convicted criminals in the colonies were sold into servitude for terms of years, as were orphans and paupers. It was also customary for purposes of apprenticeship to place children and teenagers in indentures. Massachusetts even passed a law forbidding persons to sell themselves into servitude to escape debt. Compulsory labor was probably the most common colonial labor contract. Servitude figured in the apportionment of Congressional seats among the states in the federal constitution and lasted so long as state courts were willing to imprison citizens for debt. Then the system died out, but its social effects lived on.

Such was consistent with the English and colonial background, the basic laws governing "Master and Servant," the catch-all *Statute of Artificers and Apprentices*, and the vagrancy laws, which were to some extent a thinly-veiled technique to enforce acceptance of legal wage rates. Trades unions were criminal conspiracies (until *Commonwealth v. Hunt*, 1841). Wages were commonly set and enforced by local authority in England and the colonies under legal powers. Workers could be compelled to labor in harvests, on the roads, township walls, etc.

Such restrictions upon the worker's property right in his own labor are not difficult to understand, and in fact the very title of the "Law of Master and Servant" tells us what we need to know about social origins. Who in England labored, and who did not? That such conceptions changed with time in the New World is not surprising; that they changed so slowly and still linger on in part is perhaps more so. But the colonial background was no fruitful seedbed in which a tradition of free labor contract could grow easily. And the antilabor rulings of the Supreme Court of the United States until society intervened via Congress in the Norris-

Laguardia and Wagner Acts did not encourage ideas, in law at least, of equality of rights in labor and real property. The development of a labor contract of greater bargaining equality, with fewer inhibitions on the seller, had a long road to travel. Labor was considered an inferior order of life at the beginning, and is so considered now by important segments of American society.

Such a tradition would have come over from any European nation in the 17th and 18th centuries. The difference the English origin made was that law already existed regarding labor which had to be met on its own terms by new labor facts, unions and mass production employment. Thus our history of labor law. The frustrations of a legal tradition running against the grain of American social expectations some have believed accounts for the peculiar violence of the American labor movement's history. The ultimate political conservatism of American labor, Selig Perlman argued, was due to another legacy, the tenure of real property rights. The worker could easily become an owner and that quelled any revolutionary ambitions. He would not overthrow his own.

IV. Business Practice

The exchange of property in chattel goods is the central point of commercial transactions. Anciently, such transfers lay under several constraints, the rules of established markets, *assumpsit* and contract law. Established markets (markets overt) were an ancient commonlaw device to protect both seller and buyer by the convention that titles to goods sold in such markets were exchanged before witnesses. In early colonial times one sees such market towns (and fairs) set up by statutes. The practice was inappropriate to a frontier society and by the Revolution, according to James Kent, had become obsolete, along with the special merchants' courts held in markets overt. What survives of that legacy are fairs, farmers and public markets, and blue laws. Over-the-counter sales by manufacturers, wholesalers and retailers, replaced markets overt. But other parts of the English tradition remained. *Caveat emptor* protected the seller, the ancient rules of *assumpsit* (strict liability in tort in its latest avatar) protected the buyer, and contract (formalized in part in 1677 by the *Statute of Frauds*) enforced both sides of the bargain.

Controls over the quality of goods and services vended were enforced in England in medieval times, directly through guilds, local courts and generally via *assumpsit*. In the colonial era such controls were enforced by a veritable army of viewers, searchers, wardens, gaugers *etc*. Such controls, as the Handlins and Louis Hartz showed, continued to expand in the pre-Civil War era. No part of the apparatus of municipal regulation seemed to be more ubiquitous, or more vigorously pursued in colonial times than these, judging from legislation and court records. When the enforcement of such quality controls reached the federal level (the Import Tea Act of 1897, the Pure Food Act of 1906) and flourished in our own era, ancient English and colonial restraints were being wielded.

Similarly, periodic price controls imposed in war and other emergencies in republican America were exercises of ancient powers. The assize of beer, wine and bread in *Magna Carta* was the basic power to control prices. In colonial times such powers were exercised, and even at the lowest governmental levels, as in medieval England. In the fateful decision, *Munn v. Illinois*, Chief Justice Waite referred to this long tradition to justify the Illinois law governing grain elevators. Such powers had been used:

> . . in England from time immemorial, and in this country from its colonization to regulate ferries, common carriers, hackmen, bakers, millers, wharfingers, innkeepers . . and in so doing to fix a maximum charge to be made for services rendered, accommodations furnished, and articles to be sold. To this day, statutes are to be found in many of the States upon some or all of these subjects, and we think it has never yet been successfully contended that such legislation came within the constitutional prohibitions against interference with private property.

In the cases of those public callings we usually consider public utilities, both price and

quality controls occur together now, and did in colonial times, and in England before that. Again the rule of *assumpsit* was applied; in a law of 1285 the words occur that are still used: that all who apply must be served, competently, and at a reasonable price. Originally controls of number, entry, price and quality were all applied to businesses connected with the transportation of goods and persons, inns, ferries, coaching, cartage, draymen, wharves. These sorts of businesses tended to fall between local jurisdictions, between the manor courts and those of the towns and boroughs. They were typically subjected to special legal restraints as a result. In the colonial era such controls were ubiquitous, and of course continued into the federal era and down to our own day. Those engaged in interstate commerce by rail were subject to federal regulation after 1887, however irregular its effectiveness was.

Again, the Sherman Act of 1890 was of ancient provenance. As Handler, Letwin and Thorelli showed, the Sherman Act had abundant common law precedent And when the Sherman Act used the word "every" to number the restraints of trade to be prohibited, the common law "rule of reason" had to be negotiated by the courts—apparently just because it was there. The English laws of bailment were continued in the United States after Independence, as was the bulk of English procedures regarding such problems as distraint of property for debt

The major laws relating to trade on the sea of British mercantilism were adopted in 1792 and 1793 by Congress, and became our laws for the same mercantilist reasons, to support our own maritime industry. Similarly, the protectionism of mercantilist Britain provided the new republic with the spirit of its first economic law, the protective tariff of July 4, 1789.

V. The Legacy

This is a desperately brief survey of a very large subject. But perhaps enough has been said to put flesh around the words "English legacy" regarding the institutions of economic life this country had from its beginning. Not that Americans did not create much of their own; they did. The frame of settlement beyond the 13 colonies, the Northwest Ordinances of 1785 and 1787 were themselves innovations in social thought as important as any of our organic documents. The development of the American corporations and the attendant law surrounding it owed little directly to the English. Riparian law developed here in ways unknown to Blackstone. But fundamentally the English framework, developed for centuries in English experience and law, transported here in the colonization produced an enveloping frame which was changed but not abandoned. Whatever else happened we remained identifiably English, and largely 18th century English at that, in the fundamentals of our economic institutions. The results can be astonishing, like underground oil considered in our law to be wild game. The old system was capable of unimaginable development, as subsequent American economic history showed.

There remains one further point, the absence of quantitative control in the tradition. The controls were imposed at four points in the flow of economic life: number, entry, price and quality.

These choices combine into the familiar permutations of most modern nonmarket economic control techniques. They may be arranged in a box: $\begin{array}{c|c} \text{Number} & \text{Entry} \\ \hline \text{Price} & \text{Quality} \end{array}$. If no weighting (or corruption) is allowed and the activity choices are either controlled (1) or not (0), the boxes form a little sixteen element nonmarket control matrix which is, rare quantity controls apart, essentially the nonmarket control world we live in. Barbers, for example are usually $\begin{array}{c|c} 0 & 1 \\ \hline 0 & 0 \end{array}$, any number may exist, but they must be licensed. Price and haircut quality are free of control. Taxi cabs in urban areas are usually $\begin{array}{c|c} 1 & 1 \\ \hline 1 & 0 \end{array}$, number limited, entry by license only, fares controlled but ride quality free. Restaurants are usually $\begin{array}{c|c} 0 & 1 \\ \hline 0 & 1 \end{array}$, number unlimited, entry licensed, price unlimited, quality of service controlled (health regulations usually). Zoning may, however, limit number. The

situation $\begin{array}{c|c} 1 & 1 \\ \hline 1 & 1 \end{array}$ is a regulated industry, and $\begin{array}{c|c} 0 & 0 \\ \hline 0 & 0 \end{array}$ contains the possibility, so far as law is concerned, of the conditions of perfect competition. Only in extreme emergencies were direct controls of the quantity of output allowed into the system. We now approach an era of our economic development when full-scale economic planning is advocated The tradition will make this particularly difficult if quantity control is to form the basis of our future planning system. Our practices and notions of congenial property rights have not included it characteristically, and it may prove a difficult lump to swallow, as indeed it did in the 1930's with the National Recovery Administration codes.

If it seems extraordinary to insist that economists accept the importance of our long-term history as a determinant of modern practice, it is probably because that history is unfamiliar. But we should not shun our history because we do not know it. Part of it comes from *Magna Carta*, after all, and we have no trouble, because of recent events, accepting the importance of Runnymede in contemporary political life. Originality is rare in social institutions, and that includes the institutions of economic life.

REFERENCES

J. R. T. Hughes, *Social Control in the Colonial Economy*, Charlottesville 1976.

Edward McGrady, *The History of South Carolina Under the Royal Government 1719-1776*, New York 1899.

Munn v. Illinois, 95 U.S. 113, 1877.

Robinson v. Campbell, 3 Wheaton 212, 1818.

Abbot Emerson Smith, *Colonists in Bondage: White Servitude and Convict Labor in America 1607*–1776, Chapel Hill 1947.

Adam Smith, *The Wealth of Nations*, New York 1937.

American Technology: Imported or Indigenous?

By NATHAN ROSENBERG*

This paper will consist of some variations upon a central theme, the theme of America's great resource abundance and its relationship to our technological history. I will argue, not only that much that was unique in the American technology experience was due to that abundance, but that resource abundance shaped American technology indirectly as well as directly, via routes and mechanisms which are still largely unexplored and therefore insufficiently appreciated. These routes and mechanisms take us beyond the issue of factor endowment, at least in the narrow sense. In order to appreciate these influences, one needs to examine the performance of the American economy, not just as an exercise in economic logic, but also as an historical process

The American colonies, and later the United States, must be understood as an extension of European—primarily British—culture in the unique circumstances of the North American wilderness. The earliest technologies employed by the European settlers, aside from some modest agricultural borrowings from the Indians, were those which they had acquired prior to the emigration to the New World. The mechanism of the transfer was the knowledge and the technical skills incorporated in each settler, plus the simple tools, utensils and implements which accompanied these settlers in the ships' cargoes. Throughout the entire colonial period, it should be remembered, the European technology was of a preindustrial nature. Even with the later emergence of industrial technologies, the skilled worker or artisan, as the remarkable instance of Samuel Slater reminds us, remained the vital carrier from the more advanced to the less advanced society.

The United States was a beneficiary of those extensive technological innovations in Great Britain which we now call the industrial revolution. The early American experience with industrialization, therefore, did not necessarily involve an inventive process, but more commonly the transfer of technologies which had already been developed elsewhere. However, although the early dependence upon European technology was very great, it was not total. Indeed, it could not have been so, since the differences in resource endowment and environment generated problems and presented opportunities which were necessarily outside the range of European experience European solutions and techniques were sometimes impossible and more often highly inefficient. For example, whereas forest resources were increasingly scarce and costly in Britain, they were embarrassingly abundant in America. As a matter of fact, wood and its valuable byproducts, such as potash and pearl ash, were often quite literally the waste products of the land-clearing procedures antecedent to agricultural settlement. Whereas through the eighteenth century (and earlier) the high price of timber as a fuel and as a building material provided a powerful incentive in Britain to innovate in iron production (including the substitution of mineral fuels for wood and charcoal) Americans were carefully exploring the possibilities of a wood-intensive technology. Although there was of course an extensive collection of European woodworking techniques upon which Americans could draw, Western Europe did not possess a resource endowment which would have made it worth while to explore the highly resource-intensive end of the production isoquant.

Thus, much of what distinguished the American experience was attributable to the fact that, when she commenced her industrialization in the first half of the nineteenth century, she did so with the pool of British experience upon which to draw, but also from a distinctly more favorable resource position. The direction of

*Professor of Economics, Stanford University.

technological innovation, especially in the early nineteenth century (and before) was a consequence of this circumstance. Much of it was specifically geared to the intensive exploitation of natural resources which existed in considerable abundance relative to capital and labor For example, in spite of America's late industrial start compared to Britain's, she quickly established a worldwide leadership in the design, production and exploitation of woodworking machinery. These included a whole range of machines for sawing, planing, mortising, tenoning, shaping, and boring, in addition to an entire armory of woodworking machines for more specialized purposes. It was characteristic of these machines that they were very wasteful of wood. Given the relative factor scarcities in the United States, however, such machines, which essentially substituted abundant and cheap wood for scarce and expensive labor, were admirably adapted to American needs. American lumber consumption per capita was several times as great as the corresponding levels for the United Kingdom, and rose very sharply in the early years of industrial development. In fact, American technology from the earliest times had been particularly devoted to innovations which assisted in the utilization of wood, or which reduced the cost of complementary inputs

Although America's abundance of resources led to a high degree of technological innovativeness and to the development of an essentially new technology in the case of woodworking, it is important to note that resource abundance could also operate as a conservative force. This was specifically the case with respect to the critical British innovations in the iron industry, at the heart of which was the substitution of a mineral fuel for wood fuel. Whereas the Americans were very quick to transfer some of the new industrial technologies, this was notably not the case in the iron industry, where it would be fair to say that America's abundance of wood caused her to lag a full half century behind the more "advanced" mineral-using technology of the industrial revolution. Similarly, the abundance of water power in New England caused a long delay in the American adoption of the stationary steam engine for industrial purposes. On the other hand, in uses where the steam engine uniquely facilitated the exploitation of abundant natural resources, it was adopted and exploited very quickly, as in transportation. America rapidly introduced the railroad and, even earlier, led the world in the design and construction of steamboats for her internal waterway system.

The main features of the transformation in American agriculture in the nineteenth century also bear the distinctive imprint of resource abundance. No situation comparable to American abundance had existed in any of the other societies of Western Europe from which settlers had come to America. In agriculture this took the essential form of a very high land-to-labor ratio, and this feature left a deep impression upon the direction of technological innovation in America. A major thrust of agricultural innovation under these circumstances, and one that became particularly conspicuous around the middle of the nineteenth century, were innovations which had the consequence of increasing the acreage which could be cultivated by a single farmer. This was achieved by a process of mechanization of agricultural operations. The process differed from mechanization in industry in one fundamental respect. Mechanization in agriculture in the nineteenth century involved no extensive reliance upon the new power sources which eventually played such a central role in transportation and industry. The mechanization of field operations in agriculture relied entirely upon animal power and this situation persisted until the large-scale introduction of the tractor in the 1920's. Animals supplied the power for the great mechanical innovations in nineteenth century agriculture—the steel plow (as well as Jethro Wood's earlier plow which already had replaceable cast-iron parts), the cultivator which replaced the hand-operated hoe in the corn and cotton fields, and the magnificent reaper which swept away what had earlier been a basic constraint upon grain cultivation—the seasonal variations in labor requirements which reached a sharp peak during the harvest season. Later, output per worker was further augmented by binders and

threshing machines and, eventually, by combine-harvesters. In corn cultivation, the development of the corn sheller and the corn picker were particularly valuable for their labor-saving characteristics.

American midwestern agriculture was, happily, relatively free of obstacles of topography and farm layout which retarded the mechanization of British agriculture. Furthermore, the introduction of barbed-wire fencing in the west—an area with few natural materials for fencing purposes—provided a long sought-after, cheap material for fencing purposes, and one which could be put into place with relatively small amounts of labor. It thus provided a labor-saving technique of land enclosure and, at the same time, made practicable the highly land-intensive techniques of livestock raising which became so characteristic of the American West. It should be remembered, moreover, that the westward movement itself had a significant initial labor-saving effect in American agriculture after 1860, since it involved a movement from largely forested land to nonforested land, and since the clearing of the latter lands involved a much smaller labor cost per acre than the clearing of the former In the South, the introduction of the cotton gin provided a substitute for the highly labor-intensive activity of manually removing the seeds from the cotton. In so doing it liberated cotton from the narrow coastal confinements to which the earlier cultivation of the long-stranded sea-island cotton was confined. It thus opened up for cotton cultivation an immense land area from which it was previously excluded.

Thus, the major technical innovations in nineteenth century agriculture consisted primarily of labor-saving and land-using mechanical devices, often extremely simple, which drew extensively upon the cumulating pool of technical skills and knowledge in the growing industrial sector. Such innovations had the primary effect of raising agricultural output per worker, but not the productivity of land, the abundant nineteenth century input—although one must add the qualification that, as in the cases of the cotton gin and barbed wire, some innovations were of a "triggering" nature which enabled certain lands to be put to uses or cultivated in ways which would not otherwise have been feasible. Furthermore, the commonly-observed recourse of the American farmer to soil management practices which led to declining soil fertility ("mining the soil") were essentially practices which substituted abundant land for other inputs.

So far I have argued that American technology took a direction the distinctive feature of which was its resource-intensive nature.[1] What does this have to do with those developments in the manufacturing sector, particularly those sectors which are usually regarded as defining what was so truly special about our technological history—the mass production of standardized products consisting of interchangeable component parts, and involving the use of highly-specialized machinery, a system so different from anything known in Europe that, by mid-century, it was widely referred to there as "The American System of Manufacturing"? I believe that American resource abundance had a great deal to do with American leadership in this direction, because this new technology was not only labor-saving but, *particularly in its early stages*, also resource-intensive. To the extent that this was so, it possessed an underlying economic rationale similar to the forest-based and arable-land-abundant technologies to which I have already referred. America had a very great advantage in the early search for a labor-saving technology because it was much less constrained by the resource-wasting characteristics of that technology in its early years. This is particularly apparent in the gun-making trade, properly regarded as the original home of

[1]It is interesting to note that American resource abundance shaped the scientific enterprise as well in a resource-intensive way, although the point cannot be developed here. The flora, fauna, and geology of America offered an enormous potential increment to the stock of European knowledge concerning the natural world European scientists were intensely interested in accumulating such information. Descriptive accounts and all forms of reporting on the natural environment of the New World were eagerly sought after and enthusiastically received in Europe The comparative advantage of aspiring American scientists was, thus, emphatically in the "export" of primary products—relatively unprocessed, descriptive accounts of specimens of the natural environment of the New World

mass production technology. But it was true in other areas as well that we could adopt a machine-using technology very early because American resource abundance offered the opportunity for trading off natural resource inputs for other, scarcer factors of production Thus, New England industry not only relied heavily upon water wheels but for long used wooden pitchback wheels which were, in a strictly engineering sense, highly inefficient. They were, however, comparatively cheap to build and thus allowed Americans to "waste" potential energy in order to minimize capital costs. Similarly, Americans showed a great preference for the high pressure steam engine over the low pressure steam engine which was much preferred in Britain. Although the low pressure steam engine was more efficient in its utilization of fuel, such engines, which were more expensive to construct, would again have involved a larger capital expenditure. American railroads notoriously tolerated steep gradients and sharp curvatures, and often took circuitous routes to avoid the construction of tunnels. A main consequence of such practices was to raise the fuel costs of railroad operation but, again, it made eminently good economic sense, in America, to substitute cheap natural resource inputs for more expensive capital inputs.

In numerous ways, therefore, resource abundance and labor scarcity, and the nature of industrial technology, thrust the American economy very quickly toward the capital- and resource-intensive end of the spectrum of techniques. These pressures, however, led to exploratory activities and to eventual learning experiences the outcome of which cannot be adequately summarized merely in terms of the factor-saving or factor-using biases just referred to. For they also led to new patterns of specialization and division of labor between firms—especially between the producers and the users of capital goods—as a result of which the American economy developed a degree of technological dynamism and creativity greater than existed in other industrial economies in the second half of the nineteenth century. I believe that this technological dynamism was due in large measure to the unique role played by the capital goods industries in the American industrialization process and the especially favorable conditions under which they operated. For these capital goods industries—I refer here primarily to that group involved in the forming and shaping of metals—became learning centers where metalworking skills were acquired and developed and from which such skills were eventually transferred to the production of a sequence of new products—interchangeable firearms, clocks and watches, agricultural machinery, sewing machines, typewriters, bicycles, automobiles. A key feature of industrialization is that it involved the application of certain basically similar production techniques to an ever-widening circle of final products. Moreover, the technological knowledge and competence which was gradually accumulated in this sector was directly applicable to generating cost reduction in the production of capital goods themselves. A newly-designed turret lathe or universal milling machine, or a new steel alloy permitting a lathe to remove metal at higher speeds—each of these innovations not only resulted in better machines but they also reduced the cost of producing the machines in the first place. Thus, although the initial shift to the capital-using end of the spectrum was generated by the unique pattern of American resource scarcities discussed earlier, it is by no means even obvious what the final outcome of this process was in terms of factor biases. For the capital-using path was also a path which, *eventually*, generated a much-increased capacity for capital-saving innovations. The attempt to deal with labor scarcity in a regime of natural resource abundance pushed us quickly in a direction where there turned out to be rich inventive possibilities. In turn, the skills acquired in a more capital-abundant society with an effectively organized capital goods sector provided the basis—in terms of knowledge and engineering skills and expertise—for innovations which were capital saving as well as labor saving. Indeed, most new products, after their technical characteristics became sufficiently stabilized, have passed through such a cost-reducing stage

during which capital goods producers accommodated themselves more efficiently to the large quantity production of the new product. American industry seems to have particularly excelled at these activities.

Aside from the highly visible, major inventions, capital-intensive technologies have routinely offered extensive opportunities for improvements in productivity which seem to have had no equivalent at the labor-intensive end of the spectrum. Knowledge of mechanical engineering, metallurgy and, perhaps most important of all, the kind of knowledge which comes from day-to-day contact with machine technology, provide innumerable opportunities for small improvements—minor modifications, adaptation to some special purpose use, design alterations, substitution of a superior or cheaper material—the cumulative effects of which have, historically, been very great.

It is important that these developments be seen in the actual historical sequence in which they occurred. For the fact that America began the growth of her capital goods sector not only with a strong preoccupation with standardization and interchangeability, but with some early experience with such techniques as well as a market which readily accepted standardized products, shaped the nature of the eventual outcome of the process of industrialization in some decisive ways. For the acceptance of standardization and interchangeability vastly simplified the production problems confronting the makers of machinery and provided the technical basis for cost reductions in machine making. At the same time it provided the conditions which encouraged the emergence of highly specialized machine producers as well as the transfer of specialized technical skills from one industrial use to another. Indeed, America's most significant contributions to machine tool design and operation and related processes—Thomas Blanchard's profile lathe, turret lathes, milling machines, die forging techniques, drilling and filing jigs, taps and gauges—were associated with specialized, high-speed machinery devoted to the production of standardized components of complex products.

I would like to suggest that the American experience with standardization, uniformity and interchangeability shaped the development of the capital goods sector in ways which subsequently made it an unusually effective agent for the generation and transmission of technological innovation. The extent of standardization significantly determines the amount of initiative which it was possible for capital goods producers to exercise with their customers. With extreme heterogeneity of products (as was characteristic of Britain) the role of the machine maker becomes merely passive and adaptive. His activities, of an essentially bespoke nature, are heavily constrained by the nature of his relationship to potential customers. Initiative then remains, as was the case in Britain, with the ultimate user of equipment, and it is extremely difficult to cater to his needs in a technologically creative way. American producers of machinery and their users had, at an early date, developed a far more successful network of interrelationships than had occurred in Britain. I would not want to argue that the technological factors which I have emphasized totally account for the better organizational relationships which emerged in America, but they do seem to be a critical part of that emergence. In America the relationship between machinery makers and customers provided for an interchange of information and a communication of needs to which the machinery producer gradually learned to respond in highly creative ways. At the same time, the machine user learned to rely with increasing confidence upon the judgment and the initiatives of the machine supplier—a judgment which was justified by an increasingly intimate knowledge, on the part of the supplier, of customer needs and effective ways of catering to these needs. It was, in some measure, the mutual confidence of these interfirm relationships which made it possible for machine makers to suppress customer preferences which were technically frivolous or irrelevant and, in this way, to reduce the cost of the capital equipment as well as that of the final product.

Thus, the greater freedom of the American

capital goods producers to exercise initiative, combined with their obvious financial incentive to increase the sale of their products, created a uniquely powerful and successful set of forces for the dissemination and adoption of new technologies. To a far greater extent than elsewhere, American capital goods producers engaged in successful promotional activities, simultaneously educating and persuading machinery users concerning the superiority of new techniques.

All of this seems to me to be an important part of the explanation for America's distinctive success, in the twentieth century, not so much in inventive activity, as in the ability to carry new inventive possibilities quickly to the stage of successful commercial introduction. In an economic world of increasingly complex products and processes, commercial success has turned, to a greater and greater extent, upon interfirm relationships, upon the ability to coordinate and to utilize the output and services of specialist contractors and specialist makers of components. American firms for long have excelled at integrating their own operations with those of their suppliers in a way which has enabled them to confine their own productive operations to a limited number of specialized activities, while at the same time deriving the benefits of specialized knowledge and technical expertise concentrated in other firms and industries. In Britain, by contrast, specialization of activities by firm was seldom carried as far as in the United States This was particularly so where quality control was an important consideration In this case there was usually a strong compulsion to produce the components internally rather than to develop a dependence upon often unreliable external sources. America's greater experience with the technology of standardization and interchangeability, by contrast, had reduced such problems to a more easily manageable basis. In Britain, the reliance upon specialist subcontractors was more limited than in America. Individual firms typically produced a larger proportion of their own inputs whereas American industry was much more prone to spin off specialist producers devoted to a narrow product range. These characteristics persist even today and could be documented in such "high technology" fields as aircraft, plastics, electronic capital goods, or chemical process plants.

All of this has taken us a long way from the initial condition of American resource abundance with which I started. Nevertheless, I believe that the American success in interfirm relationships, which has been so vital to our twentieth century technological dynamism, is, to a considerable extent, a product of the unique historical path which we have traversed because of our natural resource abundance. If time permitted it could be shown how American resource abundance reinforced the trends which I have discussed via its influence upon income level and the composition of demand. For resource abundance enabled us, through trade and the exploitation of resource-intensive activities, to achieve comparatively high levels of income very early in our history. More precisely, it made possible the early creation of a society with a large and substantial class of people of middle class means. (Crevecoeur had already pointed, in the late eighteenth century, to the predominance in America of people of "middling competence.") This class, particularly conspicuous in agriculture, provided much of the market for standardized, mass produced articles of simple functional design—indeed, one could trace this influence right up through Henry Ford's Model T, which strikingly resembled its predecessor, the horse and buggy, and which was originally purchased mainly by rural households. Furthermore, the abundance of land and the cheapness of food meant that food purchases constituted a lower proportion of even urban household budgets, leaving more, correspondingly, for the purchase of manufactured goods. Moreover, the rapid rate of growth of the stock of capital and the large size of the economy by the end of the nineteenth century created an extensive and growing market for capital goods which greatly strengthened the trends with which I have been concerned. But, although this is a story strongly complementary to the one I have just told, it will have to await another occasion.

Human Capital in the First 80 Years of the Republic: How Much Did America Owe the Rest of the World?

By ROBERT E. GALLMAN*

The first thing to establish is the number of foreign-born persons who entered the United States and remained therein, during the first 80 years or so of American national history. Unfortunately, the evidence on this point is imperfect.

We know that black Africans came to the United States as slaves, almost without exception, that the slave trade was officially closed in 1807, that while illegal importations continued down to the Civil War, the number of persons involved was probably relatively small, and that few American slaves were exported. Almost 19 percent of Americans were black in 1810. By 1860, the fraction had dropped to about 14 percent, reflecting the facts that black immigration had very nearly come to a halt, after 1807, while nonblack immigration had not.

So far as I can tell, from the work of Philip Curtin and Robert Fogel and Stanley Engerman, between four-tenths and six-tenths of the slaves ever imported into the territory encompassed by the United States were imported in the years after the Revolution. That group represented an increment of between one-third and one-half to the pre-existing black population. The American black population, therefore, was not simply a heritage of American colonial history. When W. E. B. Du Bois mourned the missed opportunity of emancipation at the time of the Constitutional Convention, he perceived (among other things) the impact on American society of the subsequent expansion of the slave class by importation and the increasing scope of the violence that would be required to do away with slavery, a point to which I will return.

The immigration of free men is no better (perhaps worse) recorded than the immigration of slaves, down to 1820. But the evidence suggests that it was a relatively unimportant component of the growth of the American population until the 1840's, when the great migrations of the Irish, Germans, and British began. Thus, in contrast to the migration of slaves, which took place chiefly in the first two and a half decades or so of the period between the Revolution and the Civil War, the migration of free men was concentrated in the decades of the 1840's and 1850's. In 1850, about 10 percent of the free, white population was foreign born, while by 1860, the fraction had risen to about 13 percent.

I

According to Francis Walker, first president of this association, immigration placed economic pressures on the native population, which led to the decline of the birth rate of that group. Thus immigrants did not augment the native population so much as they replaced potential members of the group. Recently, Larry Neal and Paul Uselding have adopted an extreme version of the Walker position, in an effort to answer a question similar to the one posed by the title of this paper.[1] Restricting themselves to free migration, they assume that immigrants replaced native Americans on a one-for-one basis. Since immigrants were disproportionately young adults, immigration amounted to a trade of infants for adults. Thus one can compute the "debt" of the United States to other countries, on account of im-

*University of North Carolina, Chapel Hill.

[1] I will use the essay by Neal and Uselding and a second one, by Uselding, as points of departure for my subsequent remarks. Lest these remarks appear unduly critical, let me say that had Neal and Uselding not written their papers, I would have been obliged to attempt estimates of the type they have so capably made, since such estimates clearly represent the first step in the pursuit of the subject of this paper. Since Neal and Uselding have made the estimates, I am free of this obligation and am able to devote my attention to the ways in which such estimates bear on the subject of this paper

migration, in terms of the savings in the costs of child-rearing realized by Americans. Neal and Uselding compute the savings and go one step farther, estimating the distribution of these savings between consumption and investment. In this connection they find (among other things) that by 1850, 5–10 percent of the American capital stock could be accounted for by investment flowing from the savings generated by immigration, the fraction rising to 10–20 percent by 1880.

The data used by Neal and Uselding are imperfect, as they fully recognize, and therefore their estimates are subject to a wide margin for error. This matter we can leave aside. There are other aspects of the calculation, however, that deserve our attention.

First, Neal and Uselding implicitly assume that native Americans had children simply to fill the ranks of labor and that any abridgment of the "need" to have children was a pure economic gain, measurable in terms of money savings. But that seems an untenable view of the determination of the birthrate, and one not congenial with the attitudes of the man from whom they obtained their model, Francis Walker. Walker's idea was that immigrants created economic pressures that induced native Americans to alter their habits of procreation. That suggests that there may have been welfare losses associated with the reduction of the native birth rate and that the savings in child-rearing costs cannot be regarded as pure economic gain. Certainly the means by which the Walker effect is obtained needs closer examination, a point to which I will return

Second, Neal and Uselding implicitly assume that immigrants were perfect substitutes for native Americans. But that is unlikely to have been the case. According to Lee Soltow, if one controls for age and occupation (farm-nonfarm), one finds that free, white native Americans were richer in 1860 than were the foreign-born. This may have been due to the fact that immigrants were, on average, poorer than native Americans when they set out on their journey to the United States, or it may reflect the costs of relocation to the immigrants. In either case, if we may suppose that the free Americans of 1860 were representative of the Americans whom the immigrants would have replaced, had the Walker effect been in force, then the disparity between the wealth-holdings of the native-born and foreign-born must be deducted from the savings in child-rearing costs achieved due to free immigration, if we are to obtain a net measure of America's "debt "

The fact that the free foreign born were poorer in 1860 than their free native peers may also indicate that immigrants were operating under special handicaps. Free immigrants came disproportionately to the North, where levels of literacy and health were very high by world and European standards. It may very well be that immigrants, on average, fell below these standards. We can be very nearly sure that continental and Scandanavian immigrants possessed lower levels of literacy, in English, than did native Americans, and it is very likely that Irish and British immigrants did, as well. And, finally, immigrants were subject to discrimination, which probably reduced their economic returns and the general efficiency of the economy.

Slave immigrants were widely regarded as less effective workers than native slaves, and in part this may have been associated with problems of language and the sense of isolation they produced. The point is not altogether relevant, in the present context, since the Walker effect is unlikely to have had bearing on the slave population.[2] But it is possible that the same kinds of effects may have been felt by free immigrants from non-English speaking countries.

The above considerations suggest that "savings" from immigration were less important

[2]That is, the kinds of mechanisms Walker had in mind were unlikely to be operative in the case of the native slave population. However, if the Walker effect was in force, then the importation of slaves presumably affected the birthrate of *free* native Americans and, thus, slaves "replaced" free native Americans. Since the prices paid for slave imports probably exceeded the rearing costs of free Americans, one could argue that if proper account of the slave trade were taken, the American "debt" to foreign countries, computed by Neal and Uselding, would be further reduced.

than Neal and Uselding suppose. But immigrants differed from native Americans in other respects, the economic effects of which are more problematic. For example, their tastes were different. Compare the census returns of whiskey and beer output in 1810, 1840, 1850, 1860. Those beer-drinking German and Scandanavian immigrants left their mark on the structure of American output. No doubt there are other, if less striking, examples.

Immigrants also brought new productive ideas. A good example, from an earlier period, has to do with the introduction of rice culture in South Carolina. Peter Wood claims it was accelerated, if not determined, by the skills of the slaves brought into Carolina, many of whom had grown rice in Africa. Apparently the masters didn't know the first thing about rice culture.

Immigrants also brought different social and political ideas with them. The incidence of political consciousness among free immigrants was high and these people played important roles in union organization and in radical politics

I have no way of attaching to this array of attributes a proper set of quantitative weights. Indeed, I am not sure what sign should be attached to some of them, if our purpose is, for example, to judge their effects on economic growth, narrowly defined. But it does seem to me that we should bear in mind that immigrants were different from native Americans in ways that may have been very important.

Finally, it is necessary to say that the Walker hypothesis, itself, is by no means accepted by all students of the subject. Our chairman today, Richard Easterlin, is on record in opposition to it. Easterlin points out that the mere fact that birthrates fell during the great migrations of the last six decades of the 19th century is not clear evidence that the migrations caused this result. Indeed, both rural and urban birthrates appear to have been declining from at least 1800 onward, including extended periods when immigration was limited. Furthermore, according to Easterlin, the free immigrants tended to concentrate in cities and this development probably retarded the rural-urban migration of natives. Insofar as this was true, immigration may have worked to hold the native birthrate at a higher level than it otherwise would have attained, the rural birthrate being higher than the urban birthrate.

While my inclination is to bow to Easterlin's learning, I find one feature of this argument troublesome. According to Easterlin, the course of the rural birthrate reflected the process by which land was occupied. But in the absence of immigration, native population could more readily have been supplied to urban places, reducing pressure on agricultural land and, thus, reducing the downward pressure on the rural birthrate. Furthermore, one wonders whether it is proper to ignore the effects of immigration on the native death rate, as Easterlin appears to do The cholera epidemics of 1832 and 1849, which killed thousands, are thought to have come to the United States with Irish and German immigrants, respectively. How many other deaths were the result of less spectacular diseases brought in by immigrants? And insofar as immigration raised the native death rate, immigrants could be regarded as "replacing" native Americans

Obviously, the effect of immigration on American rates of natural increase needs more study. But as matters now stand, it seems reasonable to suppose that the Walker effect was relatively weak, at least prior to 1861, and that immigrants chiefly augmented the native population. That being the case, what can we say about the economic effects of immigration?

II

One way to deal with the problem is to compute the value of the human capital implicit in the immigration stream, bearing in mind that insofar as the Walker effect was in force, this will give us an upper bound on the estimate we desire. Uselding has provided us with estimates of this type, relating to free immigrants. Perhaps the best way to summarize his results is to say that in selected benchmark years in the period 1839–59 the value of human capital imported was "one-half to three-fourths the order

of magnitude . . . of total gross physical capital formation . . . ," or between 5 and 10 percent of the gross national product.

Uselding advises us that the data are very weak and goes so far as to label his estimates "conjectural." Two aspects of the estimates seem particularly worth comment. Uselding is obliged to build up income streams from wage-rate data and assumptions with respect to the length of the work year. I feel quite sure that were these data and assumptions used in conjunction with the aggregate *U.S.* labor force data, in order to estimate labor earnings, the values obtained would exceed the labor earnings implicit in the *GNP* series Uselding uses as a standard of comparison. That is, I believe the wage and work year data operate to overstate the relative importance of the human capital represented by immigration. Additionally, Uselding is obliged to establish the skills of the immigrants from self-reported occupations at the time of immigration, whereas what he needs are data on occupations held by the foreign-born in the United States. Here again there is a possibility that the data tend to bias the estimates in an upward direction. In any case, the data problems that lie in the way of human capital estimates of a conventional kind suggest the desirability of supplementing them with other approaches to the question.

Immigrants were disproportionately young, male adults; thus the marked increase of free migration in the last two decades of the antebellum period tended to raise the fraction of the population with the highest labor force participation rate. The effect was quite surprisingly strong. For example, the fraction of the free, white population composed of the foreign born was about 10 percent, in 1850, and about 13 percent, in 1860. But Soltow reports that almost 18 percent of free, white adult males were foreign born, in 1850, and that the fraction rose to almost 26 percent, in 1860, a value not markedly exceeded in the late 19th century and early 20th century, during the flood tide of international migration. Thus, immigration increased the work force, and at a pace faster than the rate of growth of population. The effect must have been to increase the rate of growth of national product, and also, perhaps, the rate of growth of product per head.

Estimates of these effects could easily be worked out, following familiar lines of analysis. We would be bedeviled, once again, with the problem of judging the "quality" of immigrant labor, of course. But there are three other considerations that lead one to put off making the necessary calculations. First, as pointed out earlier, immigrants came to the United States with few resources. The flood of migration in the last two decades before the Civil War must have produced factor proportions quite different from those that would have been obtained in the absence of this dramatic flow and it is not altogether clear to me how we can take that into account. Second, even if immigration increased the rate of growth of per capita output—which may be true—it is by no means clear that we could conclude that immigration raised the rate of economic growth. Since the immigrant population, compared with the native population, consisted disproportionately of adults living alone, the immigrant population must have been a relatively expensive one to maintain. A rise in per capita income would be required, if pre-existing American standards were to be prevented from falling. Thus the measured rise in per capita income incident on immigration would be partly a specious rise, from the standpoints of economic growth. Finally, insofar as the *U.S.* birthrate was already falling prior to the period of heavy immigration just before the Civil War, the structure of the American population was already shifting and immigration simply hastened the transition, rather than creating it.

There is a second respect in which free immigration probably hastened the transition of the American economy. Free immigrants came disproportionately to urban areas and worked disproportionately in nonagricultural activities. The drift in the structure of the American population prior to the great surge of migration just before the Civil War was thus probably accelerated. The term "probably" is inserted to recognize the probability that immigration slowed the

structural transformation of the native population, and the slim possibility that immigrants simply replaced native Americans who would otherwise have moved to the cities and to the nonagricultural occupations. It is worth remarking, also, that insofar as immigrants settled in cities and insofar as their presence retarded the urban migration of native Americans, American society was being partitioned into at least two components, one dominated by native Americans, the other by the foreign born. This result had very wide social and political ramifications.

Free immigrants were concentrated regionally, as well. They came disproportionately to the North and insofar as they augmented the Northern population, they played a decisive role in the Civil War. Based on Soltow's data and some plausible assumptions, over 35 percent of the adult males in the North in 1860 were foreign born. These people were relatively poorer than their native peers—less able to buy out of military service—and their ideological commitment to emancipation was relatively strong. It may well be that as many as half the soldiers of the Union Army were foreign born. That was not true of the Confederate army.

Michael Shaara describes the Confederate and Union armies on the eve of Gettysburg in the following way:[3]

The Confederate army is "an army of remarkable unity, fighting for disunion. It is Anglo-Saxon and Protestant. Though there are many men who cannot read or write, they all speak English. They share common customs . . ."

The Union army is "a polyglot mass of vastly dissimilar men, fighting for union. There are strange accents and strange religions and many who do not speak English at all . . ."

That seems to me to be a good summary statement. Furthermore, the North won the war not because it had superior armaments or superior industrial power, or God knows, superior generalship, but because it was numerically far superior and because it finally found generals capable of exploiting that advantage. The foreign born made the difference.

It is a matter of more than passing interest that two sets of immigrants who came to this country between the Revolution and the Civil War played such prominent roles in the genesis and resolution of that most "American" and most important of American wars, the Civil War. The condition of the children of early slave immigrants was a chief factor in the onset of the war, while the participation of the later-arriving, free immigrants determined the outcome of the war.

It is difficult to know precisely how to fit these facts into a comprehensive and satisfactory economic analysis. But one imagines they are far more important than the facts previously elicited in the paper, that fall more easily into a conventional analysis.

[3]I am obliged to Matthew Gallman for this reference and for a helpful discussion of the point

REFERENCES

Philip D. Curtin, *The Atlantic Slave Trade, A Census*, Madison 1969.

W. E. B. Du Bois, *The Suppression of the African Slave Trade*, New York 1896.

Robert W. Fogel and Stanley L. Engerman, *Time on the Cross*, Boston 1974.

Larry Neal and Paul Uselding, "Immigration, A Neglected Source of American Economic Growth: 1790 to 1912," *Oxford Econ. Pap.*, March 1972, *24*, 68–88.

Michael Shaara, *The Killer Angels*, New York 1975.

Lee Soltow, *Men and Wealth in the United States, 1850–1870*, New Haven and London 1975.

Paul Uselding, "Conjectural Estimates of Gross Human Capital Inflow to the American Economy: 1790–1800," *Explorations in Econ. Hist.*, Fall 1971, *9*, 49–61.

Peter H. Wood, *Black Majority, Negroes in Colonial South Carolina from 1670 through the Stono Rebellion*, New York 1974.

U.S. Bureau of the Census, *Historical Statistics of the United States from Colonial Times to 1957*.

THE INVISIBLE HAND AND OTHER MATTERS

Adam Smith on Human Capital

By JOSEPH J. SPENGLER*

I shall examine Adam Smith's treatment of human capital under five heads: the optimizing system of natural liberty; the nature of human capital; its sources; its unnecessary costliness; and obstacles to its optimum use. His recourse to cost-utility criteria in assessing educational practice (134) will be touched upon but not his somewhat analytical history of educational practice over the centuries.[1]

Smith (89, 164) was aware of the past improvement in average income (89), of the elasticity of man's wants (164), and of the impact of the "gradual improvement of arts, manufactures, and commerce" (755). He did not, however, anticipate that income increase might transform educational personnel and facilities into suppliers of consumer-oriented rather than essentially producer-oriented services (164), probably because he was concentrating on then current problems.

I. The Optimizing System of Natural Liberty

Being of a Newtonian disposition to use his imagination and systematize his thought, subject to the constraints of empiricism, Smith conceptualized individual but apparently interrelated phenomena, perceiving order beneath seeming chaos in man's affairs as in nature (*Essays*, 335, 384, 392). It may be taken for granted, therefore, that Smith believed the development and use of human capital as of other resources to be closely associated with the degree to which the system of natural liberty, together with free competition, was allowed to prevail (141, 147, 343, 642, 651). Smith was cautious, however, when searching for models, "mere inventions of the imagination" (*Essays*, 384), to represent the *empirical* world of the particular He warned, for example, against translating a set of moral doctrines "into a scholastic or technical system"—"one of the most effectual expedients, perhaps, for extinguishing whatever degree of good sense there may be in any moral or metaphysical doctrine" (*MS*, 425–26).

While Smith believed man to be an empirical utility augmenter rather than a modern "utility maximizer" and the seventeenth-century invisible-hand metaphor to approximate socioeconomic reality so long as natural liberty prevailed (423; *MS*, 264–65), he was no Leibnitzian. There was often need for man's pursuit of "self-love" to be constrained by the "sacred laws of justice" (*MS*, 120–21, 224–28, 262, 249–51), presumably in the shape of an appropriate institutional matrix evolved through experience and free of the delusion that a sovereign, absolute or otherwise, could superintend the industry of private people and most effectively employ their inputs (651) Under the system of natural liberty, on the other hand, the individual in search of his own security and gain could be led, as "by an invisible hand to promote an end which was no part of his intention," namely, the welfare of other men, of society (423).

II. Human Capital and Its Sources

Smith included under "fixed capital" besides useful machines, profitable buildings, and improvements of land,

> the acquired and useful abilities of all the inhabitants or members of the society The acquisition of such talents, by the maintenance of the

*Professor of Economics Emeritus, Duke University

[1]In the text references to *Wealth of Nations* identify pages only, those to *Moral Sentiments*, pages preceded by *MS* *The Essays of Adam Smith*, London 1872, and *Lectures on Justice, Police, Revenue and Arms*, New York 1956, are referred to several times

> acquirer during his education, study, or apprenticeship, always costs a real expence, which is a capital fixed and realized, as it were, in his person. These talents, as they made a part of his fortune, so do they likewise of that of the society to which he belongs. The improved dexterity of a workman may be considered in the same light as a machine or instrument of trade which facilitates and abridges labour, and which, though it costs a certain expence, repays that expence with a profit [265–66]

The reward of human capital must reflect the investment embodied in it even as does the return on other fixed capital. "A man educated at the expence of much labour and time to any of those employments which require extraordinary dexterity and skill, may be compared to an expensive machine "

> The work which he learns to perform, it must be expected, over and above the usual wages of common labour, will replace to him the whole expence of his education, with at least the ordinary profits of an equally valuable capital It must do this too in a reasonable time, regard being had to the very uncertain duration of human life, in the same manner as the more certain duration of the machine. [101]

Degree of investment in human capital thus accounted for differences in the wages of labor as well as in the pecuniary recompence of professional people (e.g., painters, sculptors, lawyers, physicians) whose education was much more "tedious and expensive" than that of the poor man whose "patrimony . lies in his hands" (102, 122). Whence the "pecuniary recompence" of professional people "ought to be much more liberal" and was so "accordingly" (101–02, 122).

The sources of human capital were twofold: *experience* which was associated closely with the specialization of activities in an economy based on division of labor and *education* realizable mainly in schools and colleges or through arrangements such as apprenticeship. While investment in human capital turned on access to experience in specialized activities or to training under educational arrangements, it was animated by man's desire to better his condition (81, 324, 329), to improve his well being and status. Investment in human capital was complementary to that in other capital even as "tolerable security" was essential to the growth of a nation's stock of (human and nonhuman) capital and its employment (267, 268).

Innate differences contributed in very minor measure to individual differences in embodiment of human capital.

> The difference of natural talents in different men is, in reality, much less than we are aware of; and the very different genius which appears to distinguish men of different professions, when grown up to maturity, is not upon many occasions so much the cause, as the effect of the division of labour. The differences between the most dissimilar characters, between a philosopher and a common street porter, for example, seems to arise not so much from nature, as from habit, custom, and education. When they came into the world, and for the first six or eight years of their existence, they were, perhaps, very much alike, and neither their parents nor playfellows could perceive any remarkable difference About that age, or soon after, they came to be employed in very different occupations The difference of talents comes then to be taken notice of, and widens by degrees, till at last the vanity of the philosopher is willing to acknowledge scarce any resemblance. [15–16]

The capacity of division of labor to afford a range of experience productive of human capital was limited by constraints (e.g., limited availability of capital and extent of the market; deviation from the system of natural liberty as when state enterprise replaced private) on the extension of specialization (14–15, 345–48, 384, 651) and by the availability of education preventive of adverse concomitants of division of labor (737–40, 768). While general progress tended to weaken these constraints, public measures were required to prevent the adverse mental effects associated with extreme decomposition of productive activities

Smith pointed out that in rude societies necessity and limited occupational variety kept "knowledge, ingenuity, and invention" alive at simple but adequate levels. In *advanced* societies, however, complexity of the occupational structure, together with extreme specialization of skills, was purchased at the cost of mass ignorance of most matters and the threat of extinction of "the nobler parts of human charac-

ter" in the "great body of the people" (735-36). Moreover, "division of labour, having reduced all trades to very simple operations," affords "an opportunity of employing children very young," thereby greatly increasing the opportunity cost of educating the young in *advanced* England. In contrast, in *underdeveloped* Scotland the "meanest porter" could read and write (*Lectures*, 256) since there young children had little access to employment and hence went to school.

> The man whose whole life is spent in performing a few simple operations, of which the effects too are, perhaps, always the same, or very nearly the same, has no occasion to exert his understanding, or to exercise his invention in finding out expedients for removing difficulties which never occur. He naturally loses, therefore, the habit of such exertion, and generally becomes as stupid and ignorant as it is possible for a human creature to become. The torpor of his mind renders him, not only incapable of relishing or bearing a part in any rational conversation, but of conceiving any generous, noble, or tender sentiment, and consequently of forming any just judgment concerning many even of the ordinary duties of private life. Of the great and extensive interests of his country he is altogether incapable of judging; and unless very particular pains have been taken to render him otherwise, he is equally incapable of defending his country in war. The uniformity of his stationary life naturally corrupts the courage of his mind, and makes him regard with abhorrence the irregular, uncertain, and adventurous life of a soldier. It corrupts even the activity of his body, and renders him incapable of exerting his strength with vigour and perseverance, in any other employment than that to which he has been bred. His dexterity at his own particular trade seems, in this manner, to be acquired at the expence of his intellectual, social, and martial virtues. But in every improved and civilized society this is the state into which the labouring poor, that is, the great body of the people, must necessarily fall, unless government takes some pains to prevent it. [734-735]

Turning next to the apprenticeship system, Smith described its administration as out of keeping with the system of natural liberty and as having become a source of unnecessary inequalities "by not leaving things at liberty." It restricted competition by limiting the number of apprentices, and prolonging unnecessarily the period of apprenticeship, thereby facilitating price maintenance and turning the terms of trade against country people and agriculture (99-103, 118-27). Smith's concern here is with both inequalities in earnings that flow from want of perfect liberty and those issuing from the nature of employments themselves, such as their agreeableness, their steadiness, degree of trust involved, probability of success, and, as noted earlier, variation in the cost of learning different trades and types of business (Bk. I, ch. 10).

Smith's main concern, when treating of educational institutions, was the extent of their public character (Bk. V, ch. 1), together with the degree to which the cost of education, even though "beneficial to the whole society," should be borne by the state. "Were there no public institutions for education, no science would be taught for which there was not some demand, or which the circumstances of the times did not render it either necessary, or convenient, or at least fashionable, to learn" (753). However, there were forms of education beneficial to the community and yet not likely to be supplied adequately in the absence of governmental support. One instance, as we saw, was education (e.g., reading, writing, and arithmetic) at parish schools to prevent the otherwise adverse effects of division of labor; another was education suited to prevent cowardice, gross ignorance, and stupidity (738-40, 768). "In free countries, where the safety of government depends very much upon the favourable judgment which the people may form of its conduct, it must surely be of the highest importance that they should not be disposed to judge rashly or capriciously concerning it" (739-40). The requisite education "must in most cases be made up by the general contribution of the whole society" if the contributions of the immediate beneficiaries of this education (e.g., the common people in contrast with those of rank) were inadequate (768). It was desirable also that the state promote some arts and callings which, however useful, were not self-supporting (741-743).

A final obligation of the state was support of such education as was necessary to dissipate

frenzied superstition and antiscience, fundamental barriers to the growth of knowledge. The state might "render almost universal among persons of middling or more than middling rank and fortune" a knowledge of "science and philosophy" by "instituting some sort of probation, even in the higher and more difficult sciences, to be undergone by every person before he was permitted to exercise any liberal profession, or before he could be received as a candidate for any honourable office of trust or profit." Then the "inferior ranks could not be much exposed" to "enthusiasm and superstition" since "science" was the "great antidote to the person of enthusiasm and superstition" and the "superior ranks of people were secured from it" (748). It was advisable also that the state encourage gaiety, antidote to religious fanaticism, "by giving entire liberty to all those who for their own interest would attempt, without scandal or indecency, to amuse and divert the people by painting, poetry, music, dancing; by all sorts of dramatic representations and exhibitions" suited to dissipate "that melancholy and gloomy humour which is always the nurse of popular superstition and enthusiasm" (748).

III. Costliness and Suboptimal Use of Human Capital

Smith pointed to the unnecessary costliness of human capital as well as to its suboptimal use. Illustrative of excessive costliness was training under apprenticeship systems "The institution of long apprenticeships" afforded no security against bad work; it made for aversion to work since the apprentice had to wait so long to benefit (122). Moreover, long apprenticeships were "altogether unnecessary;" one could learn many of the "trades, the crafts, the mysteries" in a very much shorter time than was required under the apprenticeship system (123). Indeed, "not only the art of the farmer, the general direction of the operations of husbandry, but many inferior branches of country labour, require much more skill and experience than the greater part of mechanic trades" and yet no apprenticeship is prescribed for farming (126–27). Smith pointed to the fact that although the duration of apprenticeships in Scotland was much shorter than elsewhere, they sufficed and at the same time did not rob the poor man of the property he had in his own labor and bodily skills (121, 122).

> In every profession, the exertion of the greater part of those who exercise it, is always in proportion to the necessity they are under of making that exertion. This necessity is greatest with those to whom the emoluments of their profession are the only source from which they expect their fortune, or even their ordinary revenue and subsistence. In order to acquire this fortune, or even to get this subsistence, they must, in the course of a year, execute a certain quantity of work of a known value; and, where the competition is free, the rivalship of competitors, who are all endeavouring to justle one another out of employment, obliges every man to endeavour to execute his work with a certain degree of exactness The greatness of the objects which are to be acquired by success in some particular professions may, no doubt, sometimes animate the exertion of a few men of extraordinary spirit and ambition Great objects, however, are evidently not necessary in order to occasion the greatest exertions. Rivalship and emulation render excellency, even in mean professions, an object of ambition, and frequently occasion the very greatest exertions. Great objects, on the contrary, alone and unsupported by the necessity of application, have seldom been sufficient to occasion any considerable exertion. [717]

The arrangements under which teachers worked did not maximize incentive to good performance (733–34). School and college endowments diminished the "necessity of application" in teachers as did indulgence and caprice on the part of college and university authorities, along with compulsion of students to attend upon teachers independent of their merit (717–21). "Those parts of education ... for the teaching of which there are no public institutions are generally the best taught" (721); moreover, only the useful was taught as in womens' schools (724). There was less corruption also where teachers depended upon student fees (721–22). As evidence of the discredit into which universities had fallen, Smith pointed to the frequent replacement of attending a university by travel abroad, usually a source of dis-

sipation, lack of principle and application, and bad habits. This practice did, however, deliver a father, "at least for some time, from so disagreeable an object as that of a son unemployed, neglected, and going to ruin before his eyes" (728).

Suboptimal or excessive remuneration and/or use of the services of human capital, together with inequality of its reward, was traceable to perversions of apprenticeship, corporation privileges, poor and settlement laws, wage regulations, and subsidization of professional education (e.g., clergy, teachers, lawyers, physicians, and "that unprosperous race of men called letters") (118–42, 719). Upon removal of such privileges, subsidies, and barriers, gross-income inequalities other than those properly associated with interindividual variation in human-capital investment could disappear

IV. Conclusion

Even though as Veblen remarked, Smith was writing of an essentially handicraft economy and had little awareness of the industrial revolution on the horizon, one encounters in his work many parallels with today's discussions of education—incentive, specificity of reward, excessive educational costs, unproductive training, monopoly, destructive etatism, and so on. Situated as he was in time, place, and the course of the development of economics and economic policy, Smith did not, as do today's educators, have to worry about bringing into being an unemployable and disruptive "intellectual proletariat." He was free to concentrate on disparity between the prevailing institutional structure and an optimally competitive structure, along with means to reducing this disparity. Living as he did in a handicraft economy, he did not anticipate corporate and trade-union dinosaurism, solvent of Smith's image of exchange, incentive, and "sympathy," a counterweight to the emergence of a Hobbesian world.

Smith did not treat of utility maximization within a social and economic institution à la Gary Becker. His concern was mainly to describe and seek establishment of a politico-economic environment within which man's desire and efforts to better his condition could work optimally for himself and society whether the sought object be the adequate supply of education both useful and not needlessly expensive, or the supply of physical commodities. Fortunately, he lived in a world less ridden with externalities than today's and perhaps more given to the accommodation of feedback arrangements conducive to activity-optimization.

Smith and Ricardo: Aspects of the Nineteenth-Century Legacy

By SAMUEL HOLLANDER*

The standard of reference from which perspective the early literature is evaluated by Joseph A. Schumpeter in his *History of Economic Analysis* is the (Walrasian) general-equilibrium approach towards productive organization. Ricardian procedures, according to the Schumpeter historiography, are diametrically opposed to the spirit of general equilibrium—above all to its conception of the returns to factor services as competitively-determined prices. In dealing with the determination of the laws regulating distribution—his fundamental problem—Ricardo proceeded by arbitrarily reducing the number of variables in his model until he was left with but one variable, namely profits, to be determined as a form of residual—the difference between the given value of the marginal product of labor-and-capital and the subsistence wage rate—by the one equation of his system [1]

While it is conceded on this view that Ricardo grasped better than any predecessor the conception of economic theory as a general purpose *analytical engine*, capable of yielding results "no matter what the concrete problem that is fed into it," the specific engine devised by Ricardo constituted nonetheless a "detour" in the development of economic analysis. For Turgot, Adam Smith (in significant chapters of the *Wealth of Nations*), J. B. Say, Lauderdale, and Thomas Malthus had already achieved a considerable insight into the "correct" approach towards productive organization, namely one which encompasses distribution envisaged as the pricing of requisite and scarce services,[2] and their work had been followed during the immediate post-Ricardo period by a number of men (especially M. Longfield), who wrote "above their time."[3]

Now the "mirror image" of Schumpeter's reading of the record is to be found in Maurice Dobb's recent study. Two streams of thought—two classical traditions—relating to exchange and income distribution are discerned in this account. The first originated in Smith's cost of production theory[4] whereby competition, through the operation of supply and demand, assures that market prices gravitate towards "natural" prices defined as the sum of the unit wage, profit and rent costs, the factors paid at their "natural" rates. These natural or necessary factor payments are in turn determined by the general conditions of supply and demand for labor, capital and land. This approach "etched in lightly and suggestively by Smith" was taken further by the Longfield-Nassau Senior group, by John S. Mill, and subsequently by W. Stanley Jevons and Alfred Marshall.[5]

The new line—far from constituting a "detour," the *true* classical tradition—took the form in the first place of the replacement of Smith's value theory "to make conditions of production, and in particular quantities of labour expended in production, the basic determinant [of value] alike in capitalist and in pre-

*University of Toronto The full evidence for my contentions will appear in my forthcoming *The Economics of David Ricardo*.

[1]Schumpeter, p 568ff Schumpeter's position is much the same in its essentials as that of Frank Knight, pp 37-88

[2]Schumpeter, p 474; cf., p 568, p 673n

[3]Schumpeter, p 465 See also p 560 the rejection of the labor-quantity theory by the non-Ricardians and anti-Ricardians of the 1830's, Schumpeter contends, "shows again that the Ricardian teaching was really in the nature of a detour "

[4]The so-called "adding-up-components" version, Piero Sraffa, p xxxv

[5]Dobb, p. 44ff , p 112ff It is part of the thesis that the adding-up-components cost theory of the *Wealth of Nations* implies "the possibility of treating the sphere of exchange-relations as an 'isolated system,' " by which is meant a system of exchange relations divorced from "the conditions and circumstances of production "

capitalist society."[6] But distribution is the central issue insofar as it has *logical priority* over prices or exchange values and is thus divorced from the general pricing process.[7] Particularly important in the scheme is the weight placed upon the introduction "of a social or institutional datum in the shape of the socio-economic conditions defining the level of real wages."[8] *Given* the wage rate "the conditions of production in the industry or industries producing necessities for wage-earners played a key role in determining the ratio of profits or surplus to wages, and hence (given necessary labour-expenditures in various lines of production) relative exchange values."[9]

The formal identity between the interpretations of Schumpeter (and Knight) on the one hand, and Dobb, on the other, insofar as concerns the *content* of Ricardian theory is apparent Both emphasize the divorce of distribution and exchange—the absence of a notion of distribution as a problem in factor pricing. And both lay great stress on the conception of an exogenously determined wage rate. The notion of a dual development of economic theory is also shared. But the difference is quite clear. The Ricardian characteristics in question are treated by Schumpeter as an inexcusable lapse, a failure to appreciate the nature of economic analysis, leading to a result which lacks sense. By contrast, Dobb views the same characteristics as a matter of deliberate choice reflecting an appreciation of the nature of scientific economics.

It is a further Schumpeterian contention that Ricardianism was a flash-in-the-pan. The system not only "failed from the start to gain the assent of the majority of English economists," but by the early 1830's "Ricardianism was no longer a living force."[10] This position is much the same as that of Karl Marx, although for Schumpeter Ricardo's poor fortune was a welcome sign while for Marx the (supposed) early demise was a symptom of the degeneration of economic science. The year 1830 was for Marx the dividing line between "scientific" and "apologetic" economics.[11]

It was in fact Schumpeter's belief that Mill must be excluded from that group which constitutes Ricardo's "school."[12] This evaluation is equally characteristic of Marxist interpreters. Marx himself speaks of Mill's work as an example of the "eclectic, syncretistic compendia" which characterized the period after the collapse of "scientific" political economy in 1830.[13] Along these lines Dobb has observed of Mill: "when looking back on him from a distance one can see quite clearly that in major respects his own work was much nearer to Marshall than it was to Ricardo; and that so far as his theory of value was concerned, on the contrary to continuing and improving on Ricardo, in essentials he took his stand on the position of Smith where Ricardo had been opposing him."[14]

In his recent volume on J R. McCulloch, D. O'Brien has added his authority to the view that the central Ricardian model suffered a serious decline soon after Ricardo's death For it is the general theme of this work that while McCulloch "did much to popularize economics ... it was not Ricardo's economics that he was

[6]Dobb, p 115 See also R L. Meek, (1974, p 250)

[7]Dobb, p 35 "income-distribution (e.g , the profit-wage ratio) was a *pre*-condition of the formation of relative prices." See also pp 169, 261, 266

[8]Dobb, p. 116 Dobb takes the neoclassical approach to task for making distribution appear as "something supra-institutional and supra-historical "

[9]Dobb, p. 116.

[10]Schumpeter, p 478 See also T W Hutchison, pp 428–29.

[11]Afterword to the second German edition (1873) of *Capital*, 1965, pp 14–15, Marx's rough draft notes (written 1857) *Grundrisse*, 1973, p. 883 (See also Meek, p 54)

[12]"From Marshall's *Principles*, Ricardianism can be removed without being missed at all From Mill's *Principles*, it could be dropped without being missed very greatly" (Schumpeter, p 529).

[13]Marx, 1973, p. 883

[14]Dobb, p 122 This view is very broadly held and will be found in Mark Blaug, 1958, who devotes a chapter to the "Half-Way House of J S Mill," where we read that Mill "only succeeded in upholding an emasculated version of Ricardo's system" (p. 167) But see the position of Schwartz, pp 16–17

popularizing. . . ."[15] McCulloch, runs the argument, must be considered as foursquare in the Smithian tradition. A similar revisionist interpretation has recently been put forward regarding Thomas De Quincey.[16]

We shall be concerned with the two issues raised in the preceding account: the supposed early demise of Ricardianism and the related notion of a dual development in nineteenth-century economics entailing Ricardian procedures on the one hand and embryonic neoclassical procedures based on the *Wealth of Nations* on the other. Needless to say it is impossible to approach these matters without a rather precise specification of the Ricardian paradigm, extending beyond the general questions of doctrinal position and archetypal method. Investigation not only of the *content* but also the *origins* of Ricardo's *Principles*—particularly the process whereby Ricardo, in the spring of 1813, commenced to discern what he considered to be a number of logical errors in the Smithian position—suggests that what is *characteristically "Ricardian"* is the use of a special theory of value involving an absolute standard in the derivation of the inverse relationship between wages and profits—the famous fundamental theorem on distrubution.

In the Ricardian structure, an increase in the proportionate share of wages appears as an increase in wages, expressed in terms of the measure of value—a commodity produced with constant labor input and thus constituting a labor-embodied unit. While it may seem that the identification lacks generality since all depends upon a presumed constancy of aggregate value (in other words upon the assumption of a given total labor force) we may in fact quite accurately express the Ricardian theorem in terms of *per capita* rather than *total* wages and output: An increase in *per capita* "gold" wages necessarily implies a rise in the laborer's share in *per capita* output which is of constant "value" whatever may happen to total value. The entire Ricardian scheme is thus designed to relate the rate of return on capital to the "value" of *per capita* wages (Ricardian "real" wages)—which in effect amounts simply to the proportion of the work-day devoted to the production of wages—and variations in the rate of return to (inverse) variations in the "real" wage rate.

It follows from the basic analysis that—assuming unchanged input coefficients in the production of the monetary metal, or ruling out nominal changes in money values—wage-rate increases are noninflationary and at most generate an alteration in relative prices within limited bounds; capitalists are unable to pass on increased wage costs in the form of generally higher prices. This fundamental conception, it must be emphasized, was initially formulated as a direct challenge to received doctrine based upon Adam Smith's analysis whereby wage-rate increases *are* passed on by capitalists in the form of higher prices in manufacturing industries and lower rents in agriculture. Here lies the essence of the Ricardian contribution. The significance—indeed the objective—of Ricardo's work cannot be accurately evaluated unless placed in this historical context.

It is my primary conclusion that the Ricardian theorem on distribution—the inverse wage-profit relationship—left a firm and positive impression on the work of a number of authors normally regarded as "dissenters" *par excellence* and this despite their frequent formal criticisms of Ricardo and his followers and their declared objective to break new ground or at least to refute the merit of Ricardo's divergencies from the *Wealth of Nations*. In general terms, our investigation casts doubt upon the accuracy of J. L. Mallet's famous account of the Political Economy Club meetings of 1831 as conclusive evidence of a breakdown in Ricardo's authority.

Malthus (1823, 1824), who is commonly taken for granted to be a severe critic, accepted the substance of the inverse wage-profit

[15] O'Brien, pp. 402–03. The treatment of the invariable measure of value, which is said to be "central to Ricardo's system," we are told "never interested McCulloch at all" (146)

[16] P. W. Groenewegen, p 193.

theorem, the "simplicity and apparent obviousness" of which "do not detract from its utility." Bailey, now regarded as the head of the line of anti-Ricardians, in fact accepted the accuracy of Ricardo's position regarding the wage-profit relation *on Ricardo's use of terms*, and rejected Smith's view of the linkage between wages, profits, and prices. Our investigation also reveals a considerable degree of hostility towards Bailey's substantive work *amongst* the "dissenters," particularly Read, L. F. Cotterill and Lloyd. Cotterill's comment that there are "some Ricardians still remaining" is belied by his own description of the pure labor theory as a position which "most economists maintain." And Smith's wage-price analysis he referred to as a "very heretical doctrine," which "Mr. Ricardo has very forcibly and very successfully impugned." Merivale favorably cited McCulloch's pinpointing of the "principle defects" of the *Wealth of Nations* in value theory, namely "the abandonment of the labor theory and its replacement by a theory turning partly on labour and partly on profit, rent and wages." Ricardo, he regarded, "as the real founder of the school which at present exists in England." It seems clear that Longfield envisaged the Ricardian inverse profit-wage relationship as a valid framework for a satisfactory distribution theory (namely one built around capital efficiency in its marginal application within a demand-supply context); in any event, he was directly responsible for a formal retraction by Robert Torrens (1835, 1844) of earlier criticisms of Ricardo's position. Read (1829) and Scrope (1831, 1833) were more positively hostile but the former nonetheless insisted upon an inverse relationship between the wages of regular labor and those of management, and the latter's objections relate not so much to the inverse relationship as to precise differential effects exerted by wage changes—a standard Ricardian problem. Finally, Whately's suggestion to rename political economy "Catallactics" (the science of exchanges) does not reflect a proposed deflection of subject matter to a narrowly constrained emphasis upon "the sphere of circulation."

As for the status of the "Ricardians" themselves there seems no valid reason to place them in Smith's camp as far as concerns the theory of value and distribution. It can be shown that McCulloch, Mill and De Quincey remained loyal champions of the inverse wage-profit relation and its derivation: "Even the novice is now aware that a rise in wages would leave prices undisturbed," wrote De Quincey, in contrast to the position of "the superannuated economic systems smashed by Ricardo" which envisaged wage increases passed on in the form of higher prices.

Our investigation also demonstrates that in many cases the criticisms of the "dissenters" constitute a misunderstanding of Ricardo's position, while in others the contributions of the dissenters would not have been considered objectionable by Ricardo.[17] I have in mind particularly in this latter category the emphasis upon the relativity dimension of exchange value (Bailey); the scarcity theory of rent (Thompson 1826, Senior 1821); the insistance that the cost determination of price operates by way of demand-supply variation (Malthus); the principle of diminishing marginal utility (Lloyd, Longfield); and the abstinence approach towards interest (Longfield, Scrope, Read, Senior, 1836). But of outstanding significance is the application of market demand-supply analysis to long-run wage determination (Malthus, Longfield, Torrens, Read, Scrope, Senior). Ricardo I believe stood foursquare in this *Smithian* tradition regarding wage theory. The notion of a subsistence wage as central feature of his system, the key theme in modern representations of Ricardian economics, has been seriously exaggerated.

A demonstration of the resilience of the Ricardian distribution theorem—the broad acceptance of Ricardo's criticisms of the theory of value and distribution in the *Wealth of Nations*—and the acceptability by the Ricardians of many of the propositions of the critics suggest that it is unhelpful to think in terms of a

[17]Cf the brief comment to this effect in Jacob Viner, pp 419–20

"dual development" of economic analysis during the nineteenth century, involving on the one hand Smith, the dissenters, Mill and Marshall; and, on the other, Ricardo and Marx. We do not intend to suggest an identity of objective or of procedure between Ricardo and Marshall, but the picture is far too complex to permit any such neat categorization.

REFERENCES

S. Bailey, *A Critical Dissertation on the Nature, Measure and Causes of Value,* London 1825.

Mark Blaug, *Ricardian Economics,* New Haven 1958.

C. F. Cotterill, *An Examination of the Doctrines of Value,* London 1831.

Maurice Dobb, *Theories of Value and Distribution since Adam Smith,* Cambridge 1973.

P. W. Groenewegen, *Economic Journal,* June 1973, *83.*

T. W. Hutchison, "Some Questions About Ricardo," *Economica,* November 1952.

Frank H. Knight, *On the History and Method of Economics,* Chicago 1956

W. F. Lloyd, *Lecture on the Notion of Value,* London 1834.

M. Longfield, *Lectures on Political Economy,* Dublin 1834.

Thomas R. Malthus, *The Measure of Value Stated and Illustrated,* London 1823.

———, "Political Economy," *The Quarterly Review,* Jan. 1824, *30.*

———, *Principles of Political Economy,* 2nd (posthumous) ed., London 1836.

Karl Marx, *Capital, 1,* Moscow 1965.

———, *Grundrisse: Foundations of the Critique of Political Economy,* London 1973.

R. L. Meek, *Economics and Ideology and Other Essays,* London 1967.

———, "Value in the History of Economic Thought," *History of Political Economy,* Fall 1974, *6.*

D. O'Brien, *J. R. McCulloch: A Study in Classical Economics,* London 1970.

H. Merivale, "Senior on Political Economy," *Edinburgh Review,* October 1837.

Thomas De Quincey, *The Logic of Political Economy,* Edinburgh 1844.

S. Read, *Political Economy,* Edinburgh 1829.

Joseph A. Schumpeter, *History of Economic Analysis,* New York 1954.

P. Schwartz, *The New Political Economy of J. S. Mill,* London 1972.

G. P. Scrope, "The Political Economists," *Quarterly Review,* Jan. 1831, *87.*

———, *Principles of Political Economy,* London 1833.

Nassau W. Senior, "Report on the Stage of Agriculture," *Quarterly Review,* July 1821, *25.*

———, *An Outline of the Science of Political Economy,* London 1836.

Piero Sraffa, "Introduction," *Works and Correspondence of David Ricardo I,* Cambridge 1951.

T. P. Thompson, *The True Theory of Rent in Opposition to Mr. Ricardo and Others,* London 1826.

Robert Torrens, *Colonization of Southern Australia,* London 1835.

———, *The Budget: On Commercial and Colonial Policy,* London 1844.

Jacob Viner, *The Long View and the Short,* Glencoe 1958.

R. Whately, *Introductory Lectures on Political Economy,* 2nd ed., London and Dublin 1832.

Political Economy Club: Centenary Volume, London 1921.

A Modern Theorist's Vindication of Adam Smith

By PAUL A. SAMUELSON*

Inside every classical economist is a modern economist trying to get out. In rereading the *Wealth of Nations*, it seems to me that with a little midwifery sleight of hand, one can extract from Adam Smith a valuable model that vindicates him from criticisms of Ricardo and Marx and from the general supercilious discounting of Smith as an unoriginal theorist who is logically fuzzy and eclectically empty. My general finding, as reported in these brief literary words here today and in a companion mathematical appendix, provides a vindication of Adam Smith and serves, in my mind at least, to raise his stature as an economic theorist, both absolutely and in comparison with his predecessors and successors.

I. Views on Smith

Smith is admired for his eclectic wisdom about developing capitalism, and for his ideological defense of competitive *laissez faire* as against blundering Mercantilist interferences with the market. His analysis of the division of labor, like Allyn Young's analysis of increasing returns in the 1920's, is thought to be seminal for the understanding of change, for the Chamberlinian deviations from perfect competition, and for the young Marx's concept of *alienation* of the overspecialized worker

But there you have it. As a pure theorist, Adam Smith is written down precisely because of his fuzzy eclecticism. His natural prices and wages are thought to be merely the resultants of long-run supply and demand. His pluralistic decomposition of price and of Net National Product *(NNP)* into components of wages, land rent, and of profit is criticized as emptily tautological. After his good start with the labor theory of value, Smith is thought to have blotted his copybook by introducing *ad hoc* and not-fully-explained deductions from labor's full share by landowners and capitalist owners of stock. Even Smith's accounting decomposition of national income into value added elements of wages, rents and profits has been attacked in *Capital*, Volume 2, as involving vicious-circle reasoning. Too often theorists contrast Adam Smith to his disfavor with his brilliant predecessor, David Hume, and brilliant successor, David Ricardo.

*Massachusetts Institute of Technology.

II. The Case for Smith

My reading is otherwise.

1) Smith's value-added accounting is shown to be correct by Leontief-Sraffa modeling.

2) His pluralistic supply-and-demand analysis in terms of all three components of wages, rents, and profits is a valid and valuable anticipation of general equilibrium modeling.

3) His vision of transient growth from invention and capital accumulation, which is brought to an equilibrium end with a low rate of profit and a high total of land rent, is *isomorphic* with the model of Ricardo, Malthus, and Marx. But Smith is less guilty than these three of believing in a rigid subsistence-wage supply of labor in the short and intermediate run; so Smith's transient rise in wage rates is a credit to his model's realism, wherever it deviates in emphasis from its successors.

As a theorist, I do find things to criticize in Smith. Thus, he seems never to have known how to put net capital formation into his Net National Product concept. His exposition is 1776, not 1876 or 1976, in its vagueness. However, with careful reading, we do infer in the *Wealth of Nations* a complete and valuable theoretical model.

Finally, I omit in this brief paper discussion of

pseudo-problems that have monopolized the Smith-Ricardo literature.

Although my axioms are those of the 1776 Adam Smith, my analysis from them utilizes 1976 mathematical methods, including convenient duality theory. Today, heavy mathematics will be eschewed and reference merely made to the accompanying mathematical appendix.

III. Smith's Assumptions

i) Goods, e.g., food and clothing, are produced in a time-phased way out of land and "doses" of labor-cum-raw-materials.

ii) To arrive at net consumable outputs of goods, e.g., food and clothing, one must subtract from the gross production of each the amounts of that respective good used as input components of the various industry doses.

iii) A ration of subsistence goods per laborer, e.g., m_1 of food and m_2 of clothing, is required to produce and reproduce the population. When the worker's money wage can buy more than the subsistence vector, population grows at a positive percentage rate; when the money wage buys less than subsistence, population declines exponentially; at the subsistence wage, population is constant.

iv) Workers never save and invest. Owners of land and of raw material inputs spend their wealth on food and clothing as they will. So long as the profit rate is above some minimal subsistence rate for saving, which might be zero after allowing for stochastic losses and management expenses, nonworkers do positive saving, which is never aborted. Below that minimal profit rate, nonlaborers decumulate or dissave; at the minimal profit rate, net saving and net accumulation is zero.

v) Perfect competition prevails. Land use is auctioned off for rentals. Free entry and constant returns to scale prevail. Knowledge is, or soon becomes, general.

IV. Smith's Implications

A logician, turning his deductive crank, would deduce the following properties of Smith's system.

1) Suppose it begins in long-run equilibrium. Wages are at the subsistence level. The profit rate is minimal. Depending on the pattern of nonlaborer tastes for food and clothing, land rent will be high or low; land-intensive food price will be high or low relative to clothing price; the size of the population and of the various components of raw material inventories will be high or low depending on nonlaborer tastes; and so will depend the relative distribution of *NNP* between land owners' rent and workers' wages, to say nothing of capitalists' profits if the minimal interest rate is not zero.

Most of this Ricardo missed. Some Malthus caught. Smith denies none of this, but offers little in detail.

2) Now let there be an invention. It will be viable only if, in some industry, it raises one or more of the following: the real wage there, the real rent there, or the profit rate. Except for the singular case where its incidence happens to be solely to raise land rent everywhere, the invention must transiently raise one, or more probably both, of the profit rate and the real wage rate. This initiates population growth and capital accumulation. We are in Smith's "cheerful" transient state of growth—like England rather than China or India. But ultimately, as in China and Holland, the land fills up; the law of diminishing returns on fixed land operates.

3) The system relapses into Smith's "dull state" of equilibrium with subsistence profit rate, subsistence real wage rate, and *enhanced* land rent. In effect, Smith's system maximizes rent!

4) If inventions keep recurring, the system goes through a Brownian motion in which profit rates and real wage rates average out *above* their subsistence levels, perhaps being trendless.

5) The model captures the general behavior of economic history these last two centuries if only Smith modifies his demographic hypothesis that population explodes whenever the real wage is above an unchanged subsistence level. If the needed ration of subsistence itself grows exponentially in time, then the presumption is that (a) the real wage will oscillate around an upward-rising exponential

trend, with the labor force possibly growing slowly: (b) the profit rate will meander, averaging out positive and inducing growing capital inventories; (c) land rent will tend to rise, subject to any land-saving biases in invention and to the subtraction from its rise due to the rise in real wages; (d) once we allow for alternative ways of producing the same things and for any biases in inventions, relative wage and nonwage shares of *NNP* cannot be predicted to show any definite trends; but that does not mean that minor changes contrived in labor supply can necessarily much alter the relative wage share.

These last few propositions sound much like what Simon Kuznets reports for the laws of motion of western economies, even if Ricardo and Marx failed to come as close to them as did the *Wealth of Nations*. Hats off, I say, to Adam Smith.

6) If we add to the above model a declining supply of primary "land"—that is, declining stocks of nonreproducible natural resources, such as rich seams of metal ores and coal and exhausted geologic deposits of oil and gas—we are prepared for the Club of Rome's future.

It becomes a race between invention (spontaneous and induced) and dwindling natural resources per head: the profit rate can be expected to meander in no predictable way, the real wage to grow at a slower rate (or even to suffer a declining trend). Nonwage and nonprofit share, always so important in explaining the great historic fortunes, may possibly rise. Analysis can carry prophecy no further.

V. Verdict

It is serendipitous to be able to announce, not the Scottish verdict *unproven*, but the happy finding that Adam Smith comes through with flying colors from a modern postmortem, provided we conduct it with the modicum of charity due an early pioneer.

Mathematical Appendix

The following equations vindicate Adam Smith from the principal indictments against him, and also reveal the half-untruth present in his INVISIBLE HAND doctrine.

Productivity Assumptions

Smith assumes that any of commodities, $(q_1, \ldots, q_n)$, is produced by its industry out of its labor inputs, $(L_1, \ldots, L_n)$, its land inputs $(T_1, \ldots, T_n)$, and out of produced inputs such as raw materials (or durable equipments) purchased by the various industries: so q_j will require for its production, along with T_j and L_j, also $(q_{1j}, \ldots, q_{nj})$. Smith's production functions embodying known technology can be written as

$$(1)\quad q_j(t+1) = F_j[T_j(t), L_j(t), q_{1j}(t), \ldots, q_{nj}(t)] \quad (j = 1, \ldots, n)$$

Note the time-phasing of production in (1): inputs are needed prior to the appearance of output. In (1), T_j could be a vector of elements representing heterogeneous lands of different grades.

To arrive at net available *consumption* amounts of the ith goods, $[C_i(t)]$, one writes:

$$(2)\quad C_i(t) = q_i(t) - \sum_{j=1}^{n} q_{ij}(t) \geq 0, \qquad q_{ij}(t) \geq 0$$

Whereas a modern neoclassical economist might wish to assume that inputs can be substituted for each other in a smooth way so that $F_j[\ ]$ all have well-defined partial derivatives, a classical economist like Smith usually thought that a variable "dose" of labor-cum-raw-materials could be applied to fixed land more intensively or less intensively. So one rewrites (1) as

$$(3)\quad q_j(t+1) = F_j[T_j(t), V_j(t)]$$

$$(4)\quad V_j(t) = \text{Min}\,[L_j(t)/a_{0j}, q_{1j}(t)/a_{1j}, \ldots q_{nj}(t)/a_{nj}]$$

The a_{ij}'s are non-negative. When some a is zero, it is as if its argument is absent from the expression Min[ ].

The production functions in equation (3) are postulated to have simple properties once the scale of production goes beyond the initial levels at which the division of labor does not pay. Each $F_j[\ ]$ is concave, homogeneous-first-degree, and differentiable:

$$(5)\quad F_j[\lambda T, \lambda V] \equiv \lambda F_j[T,V]$$

$$F_j[T + \Delta T, V + \Delta V] - F_j[T,V] \geq F_j[T + 2\Delta T, V + 2\Delta V] - F_j[T + \Delta T, V + \Delta V]$$

$$\partial F_j[T, V]/\partial V > 0,$$

$$F_j[T, V] - V\partial F_j[T, V]/\partial V \geq 0$$

Finally, Smith even before Malthus and Marx believed that human labor itself had a reproduction cost at that level of *subsistence* (food, clothing, etc.) at which a family could manage to reproduce itself by mortality survival and procreation. The long-run reproduction cost of total labor, $\Sigma_1^n L_j = L$, is defined per unit of L by the nonzero column vector of needed subsistence: m_1 of q_1, m_2 of q_2, . . . , m_n of q_n:

$$(6)\qquad \mathbf{m} = [m_i] = \begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix} \geq 0$$

If the real wage exceeded the subsistence vector $\mathbf{m}$, L_t would grow; if it fell below $\mathbf{m}$, L_t would decline; at exactly $\mathbf{m}$, Smith's stationary state would prevail. Evaluating the iron ration of subsistence at its market prices, $\Sigma_1^n P_j m_j$, we compare it with the market wage, W, thereby to determine the rate of population growth. Smith's simplest Malthusian relation, I write as

$$(7)\quad (L_{t+1} - L_t)/L_t = f[1 - \sum_1^n (P_j/W)m_j]$$

$$f[0] = 0, f'[\ ] > 0, f[\ ] \geq -1$$

Clearly, when the real wage is at the subsistence level $\mathbf{m}$, population growth ceases.

Smith's Early "Rude State"

For one page, Smith does have a "labor theory of value," writing (*Wealth of Nations*, Book I, ch. 6):

> In that early and rude state of society which precedes both the accumulation of stock ["capital"] and the appropriation of [scarce] land, the proportion between the quantities of labour necessary for acquiring different objects seems to be the only circumstance which can afford any rule for exchanging them for one another.. what is usually the produce of two days' or hours' labour, should be worth double of what is usually the produce of one day's or one hour's labour. . . .
>
> In this state of things, the whole produce of labour belongs to the labourer. . . .

We can make logical, even if not historical and anthropological sense of this, by postulating that land is so abundant as to be redundant and *free*, with the ratio $\Sigma_1^n T_j/\Sigma_1^n L_j$ so great as to make land ignorable. To make inventories of raw materials and crude tools ignorable takes a greater stretch of the imagination I cut the knot by postulating that outputs and inputs are *simultaneous* rather than lagged as in equations (1) and (3).

With land redundant, so that no increase in T_j has any incremental effect on q_j output, one rewrites equations (2) and (3) in the rude state as

$$(8)\quad q_j(t) = \alpha_j V_j(t) = V_j(t), \ (j = 1, \ldots, n)$$

$$= \text{Min}\ [L_j(t)/a_{0j}, q_{1j}(t)/a_{1j}, \ldots, q_{nj}(t)/a_{nj}]$$

Here, by proper choice of dimensional units of goods or of doses, we can suppress the $[\alpha_j]$ coefficients.

Indeed, if the rude state is in exact stationary equilibrium, we can ignore all timing designations and define that exact state by the following specializations of (1)–(8):

$$(9)\quad L - \sum_1^n L_j = 0$$

$$q_i - \sum_{j=1}^n q_{ij} - m_i L = 0, \qquad (i = 1, \ldots, n)$$

By virtue of equation (3)'s definition of the

fixed components of the doses, these relations become

$$(10)\quad L - \sum_{1}^{n} a_{0j} q_j = 0$$

$$-m_i L + q_i - \sum_{1}^{n} a_{ij} q_j = 0, \quad (i = 1, \ldots n)$$

These linear equations can have a positive solution $(L, q_1, \ldots, q_n)$ only if the following technological conditions for the rude state are exactly met.

$$(11)\quad 0 = \begin{vmatrix} 1 & -a_{01} & \ldots & -a_{0n} \\ -m_1 & 1-a_{11} & \ldots & -a_{1n} \\ \cdot & \cdot & & \\ & \cdot & & \cdot \\ & \cdot & & \cdot \\ -m_n & 1-a_{n1} & \ldots & \end{vmatrix}$$

$$= \det[\mathbf{I} - a_{ij} - m_i a_{0j}]$$

$$= \det[\mathbf{I} - \mathbf{a} - \mathbf{m}\mathbf{a}_0]$$

where $\mathbf{a}_0$ is the row vector of *direct labor* requirements, $[a_{0j}]$, $\mathbf{m}$ is the column vector of subsistence requirements per worker, $[m_i]$, and $\mathbf{a}$ is the n-by-n square Leontief matrix of input-output coefficients, $[a_{ij}]$.

We now vindicate Smith's equating the competitive pricing relations of his rude state with their embodied total labor requirements (direct plus indirect). There are of course no further components of the prices of the goods, $[P_1, \ldots, P_n] = [P_j] = \mathbf{P}$, than the wage component involving the money wage, W: land rent is zero, and interest (or profit) is impossible in a world of instantaneous production.

Competition assures

$$(12)\quad \mathbf{P} = [P_j]$$

$$= [W a_{0j} + \sum_{i=1}^{n} P_i a_{ij} + 0 + 0] > 0$$

$$= W[A_{0j}] = \mathbf{W}\mathbf{A}_0$$

where

$$(13)\quad \mathbf{A}_0 = [A_{0j}] = \mathbf{a}_0[\mathbf{I} - \mathbf{a}]^{-1} > 0$$

$$= \mathbf{a}_0 + \mathbf{a}_0\mathbf{a} + \mathbf{a}_0\mathbf{a}^2 + \ldots,$$

a convergent series.

Positivity and convergence in equation (13) is guaranteed by equation (11) plus the postulate that every good must indirectly, if not directly, require some labor if it is to be a good worth talking about in the rude state.

That the real wage can just buy the iron ration of subsistence was assured by Equations (11)–(13), which imply

$$(14)\quad \mathbf{P}\mathbf{m} = W = (\mathbf{A}_0\mathbf{m})W, \qquad \mathbf{A}_0\mathbf{m} = 1$$

Incidentally, (14) tends to vindicate the empirical usefulness of Smith's notion of "labour command theory of value," as against Ricardo's semantic objections.

Stationarity of the rude states' population now follows from equation (7), which takes the form in the rude state of

$$(15)\quad (L_{t+1} - L_t)/L_t = f[1 - \mathbf{A}_0\mathbf{m}]$$

$$= f[0] = 0$$

Smith's identification of *net national product* in the rude state with wages only, or with the susbsistence consumptions of the workers, is verified:

$$(16)\qquad NNP = WL + 0 + 0$$

$$= \sum_{1}^{n} P_j C_j = (\mathbf{A}_0\mathbf{C})W$$

$$= (\mathbf{A}_0\mathbf{m})WL$$

Investment and Malthusian Growth

Smith quickly turns the page on his rude state in which the labor theory of value holds. By the division of labor or otherwise, let some set of the elements of $(\mathbf{a}_0, \mathbf{a}, \mathbf{m})$ decrease. That raises equation (11)'s determinant from zero to positive. That raises the real wage above the subsistence level. That causes population initially to grow at an endogenous positive rate, like $(1 + g)^t$. If we still keep production instanta-

neous, capital and positive profits cannot yet occur. The workers get all the fruits of the invention, and devote part of that fruit to procreation and longevity. Now

$$(17) \qquad \mathbf{Pm} < W,\ \mathbf{A_0 m} < 1,\ \mathbf{C} \geq \mathbf{mL}$$

$$(L_{t+1} - L_t)/L_t = g = f[1 - \mathbf{A_0 m}] > 0$$

$$L(t) = L(0)(1+g)^t,\ q(t) = q(0)(1+g)^t, \ldots, \quad t \geq 1$$

This initial state of exponential growth, à la Malthus (1798) and von Neumann (1932), must begin to decelerate once land becomes scarce. Eventually, workers elbow each other, trample down fields, and so forth. Land must be rationed by positive rentals, which for Smith were to go to the private appropriaters of land, selling their scarce inputs in a competitive market.

As L grows more and more relative to the fixed total of land, $\Sigma_1^n T_j = T$, positive rent income arises. Depending upon how landowners spend their rent incomes on consumption goods, and workers their surplus wages on goods, an equilibrium will emerge at each level of $(T, L, C_1, \ldots, C_n)$ for all prices $(P_1, \ldots, P_n, W, R)$. Smith's resolution of each P_j into W and R components was essentially correct, despite doubts in Marx (1885). And, even in the absence of profit and differences in time-phasing of production, Smith's solution does contradict the attempt in Ricardo (1817) to measure price ratios in terms of goods' labor contents alone.

Equilibrium Restored

At any stage of growth, for the given available technology and land, T, and for any prescribed pattern of feasible total consumption, $(C_1, \ldots, C_n)$, one can solve the planner's efficiency problem of minimizing needed total labor, L:

$$(18) \quad L = M(T; C_1, \ldots, C_n), \qquad C_i \geq 0$$

$$= \underset{T_i, V_i}{\mathrm{Min}} \sum_1^n a_{0j} V_j, \quad \text{subject to}$$

$$\sum_1^n a_{ij} V_j - F_i[T_i, V_i] + C_i \leq 0, \quad (i = 1, \ldots, n)$$

$$\sum_1^n T_j - T \leq 0,\ V_i \geq 0,\ T_i \geq 0$$

This is a standard problem in nonlinear programming, as in Kuhn and Tucker (1951). On the assumption that every good needs something of both land and variable factors, the necessary and sufficient conditions for the solution can be written down in terms of equalities involving "dual variables," or Lagrangean multipliers, or "shadow prices," which are interpretable as the non-negative price ratios $[P_1/W, \ldots, P_n/W, R/W]$, where R stands for the rental of land. (If T is a column vector of lands, R will be a row vector of rentals.) The unique conditions of equilibrium involve for scalar T,

$$(19) \quad (P_j/W)\,\partial F_j[T_j/V_j, 1]/\partial V_j = a_{0j} + \sum_1^n (P_i/W) a_{ij}$$

$$(P_j/W)\,\partial F_j[T_j/V_j, 1]/\partial T_j = R/W, \quad (j = 1, \ldots, n)$$

$$\sum_1^n a_{ij} V_j - F_i[T_i, V_i] = C_i, \quad (i = 1, \ldots, n)$$

$$\sum_1^n T_j \leq T,\ R\left(T - \sum_1^n T_j\right) = 0,\ T_j > 0$$

These are $3n + 1$ independent equations that are just sufficient to determine the $3n + 1$ unknowns of the problem: $(V_1, \ldots, V_n; T_1, \ldots, T_n; P_1/W, \ldots, P_n/W, R/W)$. But equation (19), aside from having the planner's optimality interpretation, are precisely the *competitive* equilibrium conditions under Smith's postulated production conditions.

This identifies a valid element in Smith's INVISIBLE HAND doctrine: *self-interest, under perfect conditions of competition, can organize a society's production efficiently*. (But, there need be nothing ethically optimal about the $[C_i]$ specifications and their allocations among the rich and poor, the healthy and the halt!)

We indicate Smith's resolution of the price of every good into its total wage and rent components by deriving from (18) each good's total-land-and-labor requirements. We solve for the respective pairs:

$$(20)\ L_1^* = M(T_1^*; C_1, 0, \ldots, 0) \leftrightarrow C_1 = \phi_1[T_1^*, L_1^*]$$
$$L_2^* = M(T_2^*; 0, C_2, \ldots, 0) \leftrightarrow C_2 = \phi_2[T_2^*, L_2^*]$$
$$L_n^* = M(T_n^*; 0, 0, \ldots, C_n) \leftrightarrow C_n = \phi_n[T_n^*, L_n^*]$$

These $\phi_j[\ ]$ functions give the totals of land and labor required, directly *and* indirectly, to produce a net amount of each consumption good. These Smithian functions, never before written down explicitly in quite this way, are concave and first-degree-homogeneous; if the $F_j[\ ]$ functions are smoothly differentiable, as even Ricardo assumes in his arithmetic examples, so too will be the $\phi_j[\ ]$ functions. Hence, as in Shephard (1953), they will have dual unit-cost functions

$$(21)\ \phi_j^*[R, W] = \underset{T_j, V_j}{\text{Min}} \{(RT_j + WV_j)/\phi_j[T_j, V_j]\}$$

The ϕ_j^* functions have all the concavity, homogeneity, and differentiability properties of the $\phi_j[\ ]$ functions.

So, we sustain Smith against the objection that his eclectic breakdown of prices into wage and rent components is a trivial, surface relation. We write down for Smith:

$$(22)\quad P = \phi_j^*[R, W] + 0 = R\partial\phi_j^*[R, W]/\partial R + W\partial\phi_j^*[R, W]/\partial W$$

These partial equilibrium relations are well-determined by Smith's relations of general equilibrium in equation (19).

Finally, we solve for the new Smithian steady state of zero population growth after diminishing returns has brought the post-invention wage rate down to the subsistence level: we seek the L^* root of

$$(23)\quad L = M(T; m_1L + \gamma_1, \ldots, m_nL + \gamma_n)$$

where $(\gamma_1, \ldots, \gamma_n)$ represents landowners' choice of composition of their consumption goods. As Malthus realized, the equilibrium population will be larger or smaller depending upon whether rent collectors tend to spend their incomes on goods of high or low "labor intensity." Thus, their demand for "retainers" will mean greater L^* than will their demand for food or for hunting grounds.

In long-run equilibrium states where (13) holds and the real wage is at the subsistence level, the Physiocratic Land Theory of Value holds, as described in "A Modern Treatment of the Ricardian Economy" (Samuelson 1959). Landlords are faced by a linear budget constraint in choosing their γ's, namely:

$$(24)\qquad \tau_1\gamma_1 + \ldots + \tau_n\gamma_n = T$$

where the $[\tau_j]$ coefficients involve the total "socially necessary land" involved in each C_j's production, directly and indirectly and after including the land needed to produce the needed labor's subsistence.

Realistic Time-Phasing of Production

Since output is not instantaneously producible from inputs, inventories of raw materials and of subsistence wage goods are needed for steady-state production and for growth. Smith correctly recognized that the rate at which capitalist owners of such capital goods would be willing to save in order to "accumulate" them would set a limit on the system's growth and thereby generate a positive profit rate. With land fixed, new inventions ceasing, and population growing whenever the real wage exceeds subsistence, Smith correctly saw that continued saving and accumulation—contrived by capitalists' consuming less of their current profits than is available to them—must eventually induce a falling trend in the rate of profit. Finally, at a zero profit rate (over and above stochastic average losses) or at some low positive rate below which decumulation will occur, Smith's system reaches its longest-run equilibrium.

Let r^* be Smith's long-run, low positive rate of profit at which capitalists and landowners will spend all their incomes on current consumptions. With land fixed at T, no new inventions

and no change in workers' subsistence (m_i), Smith correctly wrote his equilibrium in a tripartite breakdown of national income and each competitive price into wages, rents, and profits. His complete system becomes:

$$F_i[T_i, V_i] - \sum_1^n a_{ij}V_j = m_i L + \gamma_i, \quad (i = 1, \ldots, n) \tag{25a}$$

$$P_j \partial F_j[T_j/V_j, 1]/\partial V_j = (W a_{0j} + \sum_1^n P_i a_{ij})(1 + r^*) \tag{25b}$$

$$P_j \partial F_j[T_j/V_j, 1]/\partial T_j = R(1 + r^*), \; (j = 1, \ldots, n) \tag{25c}$$

$$\sum_1^n a_{0j}V_j = L, \; \sum_1^n T_j \leq T \tag{25d}$$

$$\sum_1^n P_j m_j = W > 0, \; V_j \geq 0, \quad T_j \geq 0, \; P_j \geq 0, \; R \geq 0 \tag{25e}$$

For r^* and **m** sufficiently small, and for T and the ratios of nonworkers' taste parameters given $[\gamma_i/\gamma_1]$, these are $3n + 3$ equations for the equal number of unknowns: n V's, n T's, n (P/W)'s, γ_1, R/W, L. A meaningful solution is guaranteed to exist by virtue of the postulated properties for $F_j[\;]$.

Independently of the (γ_i) and (m_i) parameters, there is always a factor-price-frontier tradeoff between the real wage in terms of any good, W/P_j, the real rent, R/P_j, and the profit rate, r^*:

$$W/P_j = -\psi_j[R/P_j; r^*], \; \partial\psi_j/\partial r^* > 0 \tag{26}$$
$$\partial\psi_j/\partial(R/P_j) > 0, \; \partial^2\psi_j/\partial(R/P_j)^2 \leq 0$$

For $r^* = 0$, $\psi_j[\;;0]$ is derivable from equating to unity $\phi_j^*[R/P_j, W/P_j)$. For $r^* > 0$ and all inputs used up in each single use, replacing the true $F_j[\;]$ functions by $(1 + r^*)^{-1}F_j[\;]$ will give rise to new $\phi_j[\;]$ and $\phi_j^*[\;]$ functions exactly as in equations (18) to (22). Then the fundamental factor-price frontiers defined by Smith's system can be defined by

$$\phi_j^*[R/P_j, W/P_j; 1 + r^*] = 1 \tag{27}$$

For fixed $1 + r^*$, (27) defines convex contours.

With $r^* > 0$, equation (24)'s τ's have to be marked up, but are still constants so long as the m's are constants.

Prior to the system's having settled down into its long-run, time-phased steady state, one can provide for Smith's model an *endogenous* process of growth. Recognize the nonsimultaneous character of (1), and the need for capital inventories implied by such time phasing. So long as the initial rupture from the rude state is so recent that land is still redundant and rent zero, the system can grow in an initial golden age. Its rate of balanced exponential growth and the accompanying intermediate-run rate of interest or profit will provide the endogenous roots at which the supply of saving out of capitalists' profits are just large enough to provide the inventories for widening of capital goods and the advancing of wage goods for the multiplying population. If (7)'s population-growth function $f[\;]$, is given; if (6)'s $[m_i]$ and (23)'s (γ_i) for nonlaborers are known; and finally if the fraction of profit that will be saved is a known function of the interest rate $s[r]$—then there will be an intermediate growth and profit rate, $(g^\dagger, r^\dagger)$, at which golden-age saving will equal golden-age warranted investment. Had Smith been able to write down the full conditions of this transient golden-age equilibrium, he would have anticipated Marx's expanded-reproduction tableaux of *Capital, Vol. II* and would have provided Harrod and Domar with an endogenous natural rate of growth.

Needless to say, once exponential growth runs into the constraint of scarce good land, positive rent will have to be reckoned with and recourse to ever-worse land, or ever-more-crowded best land, will imply a steadily dropping growth rate and a steady fall in the profit or the wage rate (or, most probably in both), as the post-rude *cheerful* state sinks into Smith's long-run *dull* state.

MARKET AND PLAN; PLAN AND MARKET

National Economic Planning: The U.S. Case

By RICHARD A. MUSGRAVE*

Should the United States adopt a framework of national economic planning in which to conduct its economic policy and if so, what should the plan consist of? The affirmative case has been urged by the Initiative Committee for National Economic Planning and, in somewhat diluted form, is now before Congress in the Humphrey-Hawkins bill. The bill combines amendments to the Employment Act of 1946 with new provisions, requiring the President to supplement his annual and essentially short-run Economic Report with a second document Entitled the "Full Employment and Balanced Growth Plan," this document is to deal with the longer run goals and instruments of economic policy.

While it is difficult to determine from the bill just what the plan is to consist of, it is easier to note what it is not to do. The purpose is *not* to replace our largely decentralized and market-based economic system with one of centrally determined resource allocation. While longer term, sectoral projections of the United States economy will be helpful in setting the framework, these are to be indicative at best. The current discussion thus differs sharply from the classical debate of 50 years ago when the concept of planning was interpreted in a comprehensive sense and the feasibility of efficient resource allocation under centralized socialism was at stake. Though planning emerged as feasible, the victory was incomplete An efficient solution could be secured à la Lange by letting the planners simulate a competitive market, but in the process much of the socialist baggage (e.g., the labor theory of value) had to be jettisoned in favor of marginalist economics.

Moreover, the early (when as yet untried) view of global planning as a key to social utopia came to lose much of its glamour. Capitalism, to be sure, had its defects, but six decades of socialist experience proved a fair match. While macro stability could be secured and inequality be tempered, central allocation planning developed its own inefficiencies and, more important, proved prone to impairment of civic freedom In retrospect, it is easy to see why this should have been the case. The tight and continuing powers needed for the efficient conduct of global planning cannot be reconciled with the variety and obduracy of a democratic process. Given this fact, it is surprising that the socialist tradition of economic analysis, with its sensitivity to the working of socio political forces under capitalism, should have been so oblivious to a parallel dynamics in the socialist model. As should be abundantly clear by now, the existence of "contradictions" is not a flaw unique to the design of any particular social system, but an essential feature of man's socio economic condition, not resolvable by pat formulae such as planning or free markets.

Following the pattern of West European countries, such as France, Holland or Norway, our current discussion of planning pursues a more modest and pragmatic goal. What is needed is not a comprehensive national plan in which the allocation of resources between industries and firms is determined in detail (this being precisely what the market can do rather well) but a framework for comprehensive policy making in a mixed system, designed to deal with specific areas of market failure and subject to the safeguards and vagaries of the democratic process. That such failures exist requires no lengthy demonstration and a brief accounting

*Harvard University

will suffice:

1) the market mechanism, especially in its contemporary institutional form, does not automatically meet the requirements of adequate macro performance, i.e., sustained high employment, continuing growth and an only moderate degree of inflation;

2) the market mechanism will yield an efficient allocation of resources if certain conditions are met, but it will not do so where market structures are imperfect, and where external benefits or costs arise which are not accounted for by the pricing process;

3) the market mechanism gives rise to a distribution of income which at best reflects efficient factor pricing, but even then does not carry claim to ethical sanction or social standards of fairness.

This list of shortcomings is not entered to argue that the market system is hopelessly defective, but to note that public policies are needed to complement and guide it Policy measures, to be sure, may be imperfect as well, but it does not follow (as some would conclude) that therefore the system will do better without them. The conclusion, rather, is that efforts must be made to minimize defects, and the current discussion of planning should aim at just this objective. Public policy (no less than firm or household decisions) must be planned in a coherent fashion, allowing for the interplay of diverse policy instruments in achieving specified goals. Moreover, the policy horizon must extend beyond the very short run only, so as to allow for the time path of policy effects and for changing policy needs. Current demands for a "planning framework" reflect the view that existing policy institutions do not adequately provide for a cohesive set of policies and that they are too short term in orientation. Both points are well taken, but it may be questioned whether the remedy adds up to a "national plan."

To illustrate the need for policy cohesion, consider the role of stabilization. A major lesson of recent years has been that no sharp distinction can be drawn between "macro" issues as dealt with by general policies and "micro" issues as dealt with by selective tools. The traditional macro models (Keynesian, neoclassical or monetarist) have proven inadequate in their view of stabilization as attainable by aggregate demand control alone. Notwithstanding the present respite from stagflation, structural policies will be needed if the conflict between high employment and price level stability is to be overcome in the longer run. The degree of tightness in various sectors of the labor market differs at any given level of aggregate demand, thus calling for a more sectoral view of the Phillips curve and full employment targets. To achieve it, increased emphasis on labor market policies bearing on mobility, training, selective public employment and wage subsidies is needed. Moreover, coordination of regulatory policies and of state-local finances should be given a stronger role. This will help, but it will hardly be enough. Some form of national incomes policy will have to be provided, in my view, if we are to achieve the degree of social discipline, in both labor and product markets, which is needed to reconcile the twin objectives of high employment and only modest inflation. Thus, thinking about stabilization policy must extend beyond its fiscal-monetary confines and establish a closer link with structural change.

As another illustration, consider the troublesome problem of income distribution. While the economy has done well in raising the average level of real incomes, it has failed to deal adequately with the lingering problem of poverty. Nor have redistribution policies, based on taxes and transfers, been successful in dealing with the situation. On the one side, welfare is not a satisfactory remedy for the disadvantaged; on the other, adequate support where needed is withheld lest assistance should become a subsidy to leisure. Across-the-board minimum income or even negative income tax plans fail to face up to this difficulty. To deal with it, new approaches to poverty and distributive equity need be developed, including education, employment and subsidy policies, not unrelated to

those needed in the stabilization context.

Turning to the need for a longer view, growth policy offers the most obvious case in point. After reigning for two decades as the bipartisan beauty queen of economic policy, *GNP* growth has recently fallen into disrepute. It is said to worsen the quality of life and hence to be counterproductive. As I see it, the case is not against growth but for a better measure thereof, redefined to allow for external costs. Moreover, while a large part of the world's population remains close to hunger, it seems strange for the developed world to worry about excessive production. The remedy to opulence for some decades to come is economic aid, not shorter hours. Nor am I haunted by the prediction that growth is about to cease due to resource limitation, making it necessary to plan for the advent of a "steady state" society. While the energy crisis is real and requires policy planning, the problem is one of technological adaptation and progress rather than acceptance of an absolute barrier. The task of planning, I think, is how to deal with continuing growth and how to direct it into contributing towards a more equitable society, increasing the share of low income people and fueling social mobility, rather than to prepare for the stationary state.

In addition, there is an array of specific issues, solution to which involves a significant time dimension Consider the impact of demographic change on the requirements for educational services, the effect of a decline in the birth rate on the future of the social security system, the structural problems associated with the movement of population to suburbia, or the effects which shifts in industrial location from the northeast to the southwest will have on transportation and communication systems. Capital investment, especially in infrastructure, has a long life so that efficient investment planning must account for future changes in living patterns. Thus, continuity in public policy is required, and improved information regarding possible patterns of future development may be helpful in guiding private as well as public investment decisions.

It remains to be considered how policy cohesion and allowance for a longer view can best be secured and how this is to be done under the Humphrey-Hawkins bill. The bill proposes that responsibility for the requisite data collection and analysis be vested in the Council of Economic Advisors, with the close support of the Office of Management and Budget and other executive agencies. This would involve a massive expansion of the Council's function, a function which to date has been essentially focused on the design of short-term, mostly macro oriented, stabilization policies. Clearly, a drastic increase in the Council's staff would be needed to undertake this task, accompanied perhaps by reorganization into two units, one covering short-run policy and coordination, and the other one responsible for the analysis of longer-run targets and policy designs. This would cover the issues but leave open the question whether the essentially political nature of the Council (which, subject to professional integrity, *should* be responsible for the design and defense of its Administration's policy) is equally appropriate for longer-run policy planning. If not, a case may be made for vesting this responsibility with a bipartisan body. Moreover, such a body might be made representative of both the executive and the legislative branches, thereby avoiding the duplication involved in the bill's proposal to have the staff work done on both the executive and congressional levels. Moreover, creation of a separate planning agency of this sort would bridge the lives of any one administration or Congress, thus assuring the essential continuity in policy planning and thinking.

At the same time, the setting of longer-run priorities and their implementation is not something that can or should be undertaken outside the political process. One of the merits of the Humphrey-Hawkins bill is precisely its emphasis on conducting the planning process as part of the ongoing political dialogue in the country. All views should be heard, but this should not relieve the party in power from its obligation to specify its own goals and policies. The question is whether the extensive staff work needed to

meet this obligation must be done "in house" (i.e., as part of the executive establishment) so as to assure political responsiveness, or whether these needs can be served by an "outside" planning agency, based on bipartisan and joint executive-congressional control. While the choice presents something of a dilemma, more consideration should be given to other possibilities, including a compromise solution such as the French pattern where strategic use is made of semiprivate research agencies.

The Humphrey-Hawkins bill further proposes that the President's Economic Report be supplemented each year by a "Full Employment and Balanced Growth Plan" in which "for the number of years feasible" long-term priorities and their implementation are specified. I will not deal here with the bill's designation of a 3 percent unemployment ceiling as the prime priority, but rather with the general proposal that a longer-term plan be submitted each year. The question is whether a major effort of this sort need be undertaken on an annual basis or whether, following the example of other countries (e g., France, Holland, Norway) this is better done within, say, four-year intervals. Under such an approach, which I would think preferable, each incoming President might be required to submit such a plan (which, presumably would not be unrelated to the platform on which he was elected) by the end of his first year in office, subject to amendment and updating in his subsequent Economic Reports. The four-year plan would then serve as a framework in which to conduct the Administration's economic policy and thereby demand greater attention than would an annual statement, made as a mere supplement to the Economic Report.

I am aware that this procedure would run counter to the provision of the Humphrey-Hawkins bill according to which the annual growth plan is to be presented to Congress and to be voted upon by joint resolution each year thus paralleling the procedure followed under the new budget legislation. I do not find this analogy convincing. The budget contains specific provisions needed to implement the government's expenditure program, provisions which must be legislated on each year, whereas the presentation of medium or longer-term priorities and policy approaches contained in the Growth Plan is more tentative in nature. I think that it would be sufficient for Congress to vote on such a declaration of intent once every four years. In the meantime, primary concern should be with specific acts of implementation and an accounting made of progress with the longer program.

I would thus separate quadrennial action (executive and legislative) needed to set the longer-run program from annual action on the budget. The annual intertwining of the two processes as suggested by the Humphrey-Hawkins bill seems to me too complex a procedure By linking the two parts, neither may be done justice. In particular, the functioning of the new Budget Act, one of the outstanding legislative achievements of the last decade, may be overburdened and handicapped even before it has a chance to take hold.

In concluding, a note on policy semantics. As I see it, there is no question that a coherent set of public policies is needed to complement the market. The invisible hand works, but only part of the way. It is also evident that policy thinking—goals and instruments—must extend beyond the short run into the longer future. Finally, improved institutional arrangements to meet these needs are in order. At the same time, I question whether this approach should be launched under the banner of "national planning." On the one hand it may be argued that planning is but a synonym for rational policy design and thus beyond objection. On the other, the concept of national planning (as distinct from, say, the more limited use of the term in city planning) has traditionally referred to a comprehensive set of centrally prescribed allocation measures, an approach which is neither needed nor desired. Given this semantic tradition, it seems unwise to prejudice the case for consistent policy design (or if you wish, policy planning) with the loaded overtones of a call to a "national plan."

Where does this leave the ideological content of the planning versus market issue? I do not share the view that ideology does not enter, since people may plan for whatever society they wish to have. Planning and market are not the same thing. A planning or public policy regime involves social cooperation by explicit implementation of agreed-upon goals. It thus differs from an invisible hand, or market regime, where a benign social outcome results without the explicit intent of the actors. We then pose two questions, the first scientific and the second ethical. The scientific question is whether the objective factors which condition social and economic processes permit an optimum to be approached by using one or the other regime only, or whether some mix will yield superior results. The ethical issue is whether the answer to the scientific question is appealing or distasteful. *A*'s Utopia might let each person pursue his own goals with harmony established via an invisible hand, while *B* looks for one where harmony drives from explicit concensus, cooperation and social design. Even economists, I suspect, might be classified into *A* and *B* types, but as social scientists they can hardly avoid the conclusion that the objective factors call for a mixed regime, involving both market and planning (i.e., public policy) components.

The Case of Yugoslavia

By DEBORAH D. MILENKOVITCH*

In the 1960's the Yugoslav economy stood out among Marxist systems because of its heavy reliance on the profit motive and the market mechanism. In the 1970's, in developments little noted to date, the Yugoslavs have significantly altered their attitudes towards plan and market. The recent shift in attitudes is the subject of this paper. The principal conclusion is that the constitutional reforms of the 1970's aim at establishing an economic system which is neither plan nor market as these terms have customarily been used.

I. The Economic Reforms of 1965

The economic reforms of the 1950's decentralized production decisions, introduced workers councils to manage the enterprises, and activated the market mechanism, while retaining central control over the level of investment and its allocation among sectors and regions The introduction of market elements appeared to generate pressures to extend the domain of the market, much as Paul Sweezy has suggested, and after a long and bitter struggle during the years 1959–66, the controversial economic reforms of 1965 were implemented.

These reforms had three basic goals:[1] to establish self-management, to improve the tense relations among Yugoslavia's several nationalities, [2] and to attain socialism. Self-management was defined more as the absence of intervention by government agencies or by political organizations in the affairs of the enterprises than as the positive participation of workers in the decision-making process. Attainment of socialism was taken to require rapid growth and modernization, for which economic efficiency at the micro level was held an indispensable means, and a minimal role for the state bureaucracy, which the Yugoslavs saw as the chief obstacle to attaining self-managing socialism.

These goals were to be accomplished through a variety of means. Political tensions about interregional resource transfers would be diminished by reducing the volume of such transfers, in particular the volume of federal investment funds, thereby reducing the level of political consensus necessary. Enterprise taxes would be lowered, allowing higher retained earnings which would be allocated among competing investment projects by the "neutral" mechanism of the market. Enterprises were permitted to lend money directly to other enterprises or through the banks. Banks were established as financial intermediaries which, like other enterprises, were engaged in the quest for profits. It was hoped that profit-motivated enterprises and banks would allocate investment funds more efficiently than had governments pressured by special interest groups. Efficiency was also to be increased by sharpening economic incentives. Reduced taxes would make income depend more closely on enterprise performance. Markets were to be made more effective by freeing prices from controls and by liberalizing imports to compete with domestic products. The plan ceased to be a directive and became a set of social guidelines whose function was unclear.

By 1970, the economic reforms had failed in all of their objectives. (1) Self-management did not increase as much as had been anticipated. Enterprise autonomy was limited because government intervention, although diminished, did

*Professor of Economics, Barnard College and Research Associate, Research Institute on International Change, Columbia University I have benefitted from the helpful comments of Bettina Berch, Joel Dirlam, Duncan Foley, Helen Kramer, Robbin Laird, Egon Neuberger, Laura Randall and especially Stephen Sacks Remaining errors are my responsibility

[1]The 1965 economic reforms, which actually involved a series of measures introduced over the period 1961–67, had other objectives such as curbing the rate of inflation and altering the international trade and foreign exchange system which will not be treated here For a more detailed discussion of the reforms, see Milenkovitch and Horvat

[2]For brevity, I shall refer below to the republics, although the proper phrase would be the republics and the autonomous provinces, whose borders coincide approximately with ethnic groupings.

not disappear. Enterprises became highly dependent on the banking system because of high levels of indebtedness, recurring liquidity crises, and limited sources of investment funds relative to demand Within the enterprise, workers had relatively little influence over decision compared to the managerial elite. (2) While efficiency may have increased within the enterprise as a result of sharpened incentives, government intervention at the federal, republican and communal levels seemed to offset such gains. Republics (and communes) protected their own enterprises, built their own prestige projects, used political pressure to merge failing enterprises with successful ones, and raised substantial barriers to interregional capital mobility. Intermittently controlled prices continued to cause structural disproportions which in turn aggravated balance of payments difficulties and occasionally caused shortages. There was no countercyclical fiscal policy and the burden of controlling inflation fell on monetary policy. (3) The increasing economic inequality that accompanied the reforms seemed both inconsistent with socialism and politically divisive. The concentration of economic power in the hands of financial institutions, the managerial elite, and the foreign, wholesale and retail trading enterprises was uncomfortably reminiscent of capitalist economies and troublesome for a socialist state. (4) Most important, economic reforms appear to have exacerbated tensions among the various nationalities. Regional income inequality increased after the reform. The differences in levels of economic development among the regions caused the impact of public policy to differ substantially among regions, and Yugoslav republics were no more able to agree on a rate of exchange or on a monetary policy than they had been able to agree on investment policy. Economic policy was effectively paralyzed during the second half of 1970 over such issues. Price controls, which were never fully lifted, prevented markets from clearing in foreign exchange, in investment credits and in some raw materials, and there was perpetual squabbling about the interrepublican allocation of supply. During 1971 the interrepublican disagreements over economic issues grew explosive.

Within five years of their adoption, the 1965 economic reforms had failed. Whether this is because the profit motive and the market mechanism are inherently incompatible with socialism; whether it is because the market mechanism in the particular Yugoslav environment, with large differences in levels of development among regions and high interrepublican tension, was bound to be politically divisive; or whether it is because the reforms were implemented in a disorderly way cannot be further explored here. What can be said is that the reforms not only failed, but the intensity of the interrepublican rivalries which were in part the product of the reforms, appeared in 1970–71 to threaten Yugoslavia's continued existence as a state through possible moves for secession or, more likely, through internal political disarray which would invite Great Power intervention. The leadership responded to the crisis by introducing a series of major constitutional changes which not only altered relations between republics and the federal government, but which also provided for major reforms in the system of economic organization

II. The 1971 Amendments and the 1974 Constitution

Proposals for constitutional changes were first made in the fall of 1970, and in June 1971 the Federal Assembly passed 23 amendments to the 1963 constitution. In February 1974 the Federal Assembly adopted an entire new constitution; its philosophy, however, was the same as that of the 1971 amendments.

The aims of the reforms in the system were essentially the same as in 1965—to resolve the nationalities question, to establish genuine self-management, to attain socialism—but these objectives were now defined quite differently.

A. National Equality

Whereas the purpose of the 1965 reforms had been to minimize areas of potential conflict, by the 1970's focus shifted to defining procedures for reaching consensus and for dealing with

stalemates. The constitution gave the federation only specifically enumerated functions. These functions were divided into two groups: those where the federal legislative and executive organs could act independently of the republics (defense, foreign affairs, economic relations with foreign countries, and protection of the socialist system); and those which required prior consultation with the republics (monetary policy, foreign trade and foreign exchange, price control, the social plan, aid to less developed regions, the rate of the turnover tax, and the federal budget). In the event that the republics failed to reach agreement, procedures were defined for emergency legislation.

B. Self-management

Whereas in 1965 self-management was taken to mean primarily enterprise autonomy, the 1970's definition emphasized the participation of workers in the decision-making process and the right and obligation of the workers to allocate the income produced in their own production unit. Implicitly, the principal Yugoslav criterion of socialism was direct control by the workers of the distribution of surplus value.

The primary decision-making unit was defined as the Basic Organization of Associated Labor *(BOAL)*. In essence, any part of the enterprise whose results can be valued, either on the market or independently, should be constituted as a *BOAL*. The *BOAL*s would be relatively small, allowing workers actually to participate in decisions. *BOAL*s are independent, are free to associate with (or dissociate from) other *BOAL*s, and to form associations of *BOAL*s. Relations among *BOAL*s are governed by voluntary contracts called self-managing agreements. Exactly how *BOAL*s will coordinate their activities remains unclear. The contracts could presumably range from quite precise short-term commitments to long-term joint ventures. Associations of *BOAL*s could delegate certain powers to central administrative and central self-management bodies of the association, but decisions involving the distribution of personal income, investment and collective consumption are, by constitutional law, subject to ratification by the individual *BOAL*.

Self-management was extended further in the social services sector. Self-managing communities of interest were to be formed in education, science, culture and health. The assembly of such a community includes those working in the respective fields, representatives of the appropriate levels of government, citizens groups, trade union representatives, and representatives of the associations of *BOAL*s which by their contributions (taxes) finance these services. Communities of interest could also be developed in the production of marketed goods, as was the case for electric power in Slovenia.

As a final part of the redefinition of self-management, the active involvement of political organizations in the economic affairs of the *BOAL*s was legitimized. In particular the trade union organization was to participate actively in the business decisions of the enterprise, formulate criteria for the realization and distribution of income, and delegate members to the commissions who nominate the individual executives or members of the business boards of the enterprises (formerly the independent decision of the self-management organs of the enterprise)

C. Attainment of Socialism

Whereas in 1965 the emphasis was on efficiency and nonintervention, in the 1970's attention had shifted back to some of the more traditional features of socialism: income distribution, the role of planning, solidarity of the working class, and the leading role of the party.

There was a change in the definition of socialist income distribution. In 1965 the definition had been "to each according to his own labor," in effect as measured on the (imperfect) market and including a share of the implicit return to the other factors of production at the disposal of the enterprise. The 1970's definition became "to each according to his present labor and that portion of his *past labor* which was reinvested in the enterprise," in effect as measured on the market or by other means, and as corrected by social compacts or by statute. The explicit recognition of a return to past labor is

novel for a Marxist system. Income arising from the existence of monopoly, from uncollected differential rent and from windfall gains could no longer be distributed to personal income. Social compacts—agreements, among *BOALs*, trade unions and other organizations—established guidelines on income distribution. Because of the recognition that intersectoral distribution of income depends in part on the relative prices in various sectors, price intervention through social compacts on price policy was accepted as a legitimate means of affecting income distribution.[3]

Whereas the 1965 reforms emphasized competition, the 1970's system placed more emphasis on the solidarity of the working class. Income distribution was to depend not solely on the earnings of the individual enterprise or *BOAL* but also on average increases in the productivity of labor Social compacts established a solidarity fund, to which *BOALs* contributed, to assist enterprises in temporary economic distress. It was assumed that the solidarity of the working class would enable the individual self-managing organizations to reach agreement on distributive issues (in social compacts on income and price policy, in self-managing agreements on transfer prices between *BOALs*), and that the participants would not view all choices solely from a private vantage point. Ensuring that the broader social viewpoint receives adequate consideration is one of the principal functions of political organizations

D. Plan and Market in the 1970's

Probably the greatest difference between the 1965 and the 1970's thinking was in the roles assigned to the integrative mechanisms in the economy. Under the 1965 reforms the market mechanism had been regarded as neutral and efficient and had been intended to be the dominant integrative mechanism. Now it was regarded not only as producing illegitimate income differentials but also as inefficient. Gross duplications and structural deficiencies had occurred which the Yugoslavs tended to associate with the absence of *ex ante* coordination rather than with erroneous pricing policies or faulty incentives. It was agreed that the role of planning had to increase. In accordance with the principles of self-management, the social plan was to be based upon plans originating from below, from the self-managing organizations. These were to be drawn up using a common methodology, common assumptions, and a common planning period. They would be aggregated, reconciled for macroeconomic and intersectoral consistency, and returned to the self-managing organizations for their approval, presumably as signatories to a social compact representing part of the plan. The mechanisms through which this coordination was to be accomplished were left agonizingly vague in the 1976 Law on Social Planning.

In short, although the market mechanism could not yet be dispensed with, it was no longer sacrosanct. Further, the new interpretation of the role of the market asserted that it was a transitional mechanism to be transcended. It was to be superseded by the system of social agreements. In self-managing communities of interest, the parties directly agree upon the quantity and quality of services and the price to be paid for them. With the development of modern technology and the assumed integration of *BOALs*—especially vertically integrated complex organizations "from the furrow to the breadbasket"—transfer prices arranged by contract among the associated *BOALs* would replace market transactions.[4] With time, society, through its self-managing organizations, would increasingly determine directly what is to be produced and would determine directly the compensation to be received. In such a self-managed system the integrative mechanism would be neither plan nor market as these terms have customarily been used.

[3]The effect of price intervention on economic efficiency was rarely noted.

[4]It is by no means clear that modern technology necessarily or uniformly increases the advantages of vertical integration However, see Oliver Williamson on the factors affecting the choice among integration, contractual supply or market transactions

III. Implications of the New System

The amendments altering the economic system were first passed in 1971 and reaffirmed in the 1974 constitution. Before the system can be fully implemented, some twenty additional laws have to be passed. The first of these, the Law on Planning, was passed in February 1976. The Law on Basic Organizations of Associated Labor was under discussion in the Federal Assembly in mid-1976. Kiro Gligorov, President of the Federal Assembly, stated in an interview in May 1976 that the government hoped the remaining laws would be passed before the end of 1976. There is reputedly controversy about some of these laws (the laws on the banking and credit system, the monetary system, the foreign trade and exchange system), and it remains possible that the system will not be implemented. However, strong elements in the political leadership are committed to this system and over a period of five years have brought it to the point where it may be about to become operational. The implications discussed below are those that might follow if the system were put into effect in the form in which it has been discussed to date.

A. Changes in the Locus of Power

The constitutional changes will alter the distribution of authority in society which will lead to changes in the locus of power.[1] The reforms appear to reduce the importance of the managerial and financial elite in several ways. The establishment of *BOAL*s with substantial self-managing rights will probably weaken the central administration of the enterprises. Further, the power of the financial elite should diminish under new provisions restricting the activities of banks. Finally, the operational freedom of the foreign, wholesale and retail trading enterprises is to be curtailed under federal statutes.[2] The constitutional changes would seem to reduce the powers of the federal bureaucracy while increasing those of the republican bureaucracy. However, the republics will lose some of their power to influence the banks and may lose power to protect regional markets, so the net effect at the republican level is unclear.[3] The reforms increase self-management at the level of the *BOAL* which should increase the power of the workers. However, this may be offset by the increasingly active roles of the party and the trade unions, so again the net effect is unclear.[4] The only decisive gainers are the political organizations—the League of Communists (the party), the trade union organization and the Socialist Alliance (the mass political organization). These bodies have new explicit functions and probably new implicit functions. They are well-organized and thus able to exercise their new powers effectively. The party itself, since late 1971 (after the amendments were passed), has been purged of liberals, nationalists and centralists, has reaffirmed the policy of democratic centralism and has reasserted its own leading role in the construction of socialism. Most observers believe that what a few years ago was in effect eight different regional parties has at least temporarily been forged into a unified organization.

The shifts in the distribution of power implied by the 1970's reforms raise many interesting questions. The principal puzzle is how the allegedly powerful technical, managerial and financial interests, victors in the struggle for the 1965 reforms and gainers in the redistribution of power after that reform, so readily acceded to a reform that apparently diminished their powers and how they were so easily displaced by a party that at the time was deeply divided on both regional and ideological grounds.

B. Economic Implications

It is a bit difficult at this point to say how the system will work in reality. However, on the basis of laws passed and the general tenor of discussion about the intent of the system, it is possible to indicate in theory some of its interesting potential advantages.[5] (1) A system of planning from below would be more than just indicative planning because after the recon-

[5]The emphasis on potential advantages does not imply that there may not be significant disadvantages. A thorough analysis of the proposed system is beyond the scope of this paper.

ciling had been completed, the producing units would presumably commit themselves to output, investment, wage and price levels in social compacts, thus reducing the degree of uncertainty that is inherent in indicative planning. Such a system would have the benefits of conscious setting of social goals and *ex ante* coordination without the attendant concentration of power in the hands of the bureaucracy that is associated with central planning. (2) The system of basic organizations of associated labor, linked through a set of voluntary contracts with other *BOAL*s would have the advantages of a more human scale of operation within each unit, permitting participation, solidarity, and emphasis on common needs and interests rather than on hierarchy and competition. (3) The system of self-managing communities of interest in the social services sector would make that sector more responsive to the needs of consumers and workers rather than to the dictates of government bureaucracy (4) The relations among individuals would not be those of commodities bought and sold on the labor market, but those of solidarity, of belonging to a group. The market economy's competitive society of atomistic units would be replaced by a society in which individuals form communities in the work place, in the political or trade union organization, and in the place of residence, and are incorporated into the system on the basis of their multiple functions as consumers, citizens and workers.

Whether such a system is at all feasible; whether it is a system in which a relatively high price in terms of goods and services foregone is paid for other social values; or whether it is a system which, because of its solidarity, security, participation and sense of community, will be even more efficient than either plan or market in producing goods and services that are really desired, is simply unknown at this point. But given the potential human benefits, it is clearly a system that merits further serious consideration.

REFERENCES

Kiro Gligorov, interview in NIN (Belgrade), May 2, 1976, 5–7.

Branko Horvat, "Yugoslav Economic Policy in the Post-War Period," *Amer. Econ Rev.*, June 1971, *61*, Supplement, 71–169.

Deborah Milenkovitch, *Plan and Market in Yugoslav Economic Thought*, New Haven 1971.

Paul Sweezy, "On the Transition Between Capitalism and Socialism," in Paul Sweezy and Charles Bettelheim, *On the Transition To Socialism*, New York 1971, 15–34.

Oliver Williamson, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York 1975.

The Soviet Case

By ARON KATSENELINBOIGEN AND HERBERT S. LEVINE*

Due to the narrow confines of this paper, we will not attempt an in-depth explanation of the issues involved in the plan-market relationship in the Soviet economy. Instead, we will approach the problem by first sketching out a matrix mapping of the interactions among the three major groups of economic units found in the Soviet economy, and then discussing the different relationships between plan and market that are observed in the individual cells of this interactions matrix. In passing, we will suggest a number of issues which would warrant deeper discussion in a fuller treatment of the plan-market question

To begin, the three major groups of economic units are: state enterprises, collectives, households. To the first group of state enterprises belong all the producing units which are classified as being state property This includes all the state industrial, construction, transport, communications, and trade enterprises which are subordinate to the various economic ministries. It also includes the state farms in the agricultural sector. To the second group of collectives, belong primarily the collective farms. But the group also includes a very small number of producer cooperatives which still remain in light industry.[1] The third group is that of the households, the individual workers and consumers in the economy.

In Section I, we present a 3 × 3 interactions matrix, in which the rows represent the groups that produce and/or send goods or services, and the columns represent the groups that receive and use the goods and services. Within each of the nine cells, we list the type of interactions which can be observed in the Soviet economy. In the next section of the paper, we discuss the plan-market interactions, cell by cell. In the final section, we present some brief summary observations and conclusions.

I

Interactions among units in an economic system fall into two main categories: vertical and and horizontal. Vertical interactions are those between units in a hierarchy wherein each unit possesses administrative authority over the units below it. Horizontal interactions are characterized by direct interrelations between units, in which neither unit has administrative authority over the other. There is not, however, one type of vertical interactions mechanism and one type of horizontal mechanism. Rather, there is a broad spectrum of vertical and horizontal mechanisms. Vertical relations can take different forms, varying in the degree of control, exercized by superior units and discretionary behavior exercized by subordinate units. Horizontal relations can also be organized in different ways, varying in the degree to which the direct interaction between participating units is controlled by the center or free from control by the center.

In the literature on Soviet economic planning and centralized economic planning in general, it would appear that many hold the view that only vertical mechanisms and interactions matter, for it is there that the directions of resource allocation are set and the coordination of economic units established; and that the direct horizontal interactions themselves are of secondary, merely technical importance. Without prejudging the relative importance of vertical and horizontal mechanisms, it is our view that by focusing attention on the diverse horizontal interactions in the Soviet economy, some of the

*University of Pennsylvania and University of Pennsylvania and the Stanford Research Institute, respectively We wish to thank the Ford Foundation for its research support, and the members of the University of Pennsylvania Economic Planning Workshop for their comments on a draft of this paper

[1]For the purposes of this paper, we treat the cooperative stores, which exist in rural areas, as state enterprises, since they generally operate in the same manner as official state stores (including, in reality, the payment of wages by the state to the store employees).

richness and complexity of the plan-market issue can be brought to light.

The horizontal relations that we will discuss take three forms: 1) rationing, 2) monetary payment, 3) free acquisition (*besplatnyi*). By the term "rationing," we mean the complete vertical control of horizontal transactions. This could include both the central assignment of inputs (labor, land, plant and equipment, and materials), to, and between, particular producing units, and the central assignment of consumers goods to individual consumers. The second term denotes a transaction involving a monetary payment and the absence of direct rationing (those transactions involving both rationing and monetary payment are included under rationing). The third term refers to transactions involving "free" goods and services.

There is, again, a tendency in analysis of the Soviet economic system to associate rationing with central planning, monetary transactions with the market, and free goods with full communism. Although there is much substance in these associations, they are vastly oversimplified, hiding the great extent of diversity that exists within these classes and their relations to plan and market

One final group of categories before proceeding The sets of interactions, observed in the Soviet economy, which loosely and without striving for precise definitions may be referred to as market relations, vary in a number of ways including the degree of their legality. There are not just legal markets and illegal, black markets, but an entire spectrum of multicolored markets. Following a scheme, described by Katsenelinboigen (in a forthcoming article in *Soviet Studies*), we will use colors from bright to dark to describe the range of markets from legal to illegal:

1) Legal Markets
   Red—Prices established centrally
   Pink—Participants in transactions have some freedom to alter prices
   White—Participants set prices
2) Semilegal Markets
   Gray—Transactions illegal, but tolerated by the authorities
3) Illegal Markets
   Brown—Transactions illegal, but penalty less severe than criminal prosecution
   Black—Transactions illegal, penalty is criminal prosecution

In Figure 1, we present a matrix of horizontal interactions in the Soviet economy (columns receive goods and services from rows). We now proceed to discuss the interactions delineated in the matrix, cell by cell.

Cell #1 (State Enterprises to State Enterprises)

The interactions between state enterprises are generally viewed as the key set of relations in the Soviet command economy These relations take the following forms:

A. *Rationing*. This is the centrally planned and controlled material supply system, in which a user state enterprise must have a rationing document in order to acquire input materials produced by other state enterprises. Yet even in this core element of Soviet command mechanisms, there are elements of discretion in horizontal relations. The indicators in Soviet plans are to a substantial extent aggregated with respect to product and time detail. The participants in these horizontal interactions have the discretion to decide this detail, within the control aggregates established in the plan. Furthermore, there is also in this centralized, rationing mechanism an embryonic market element in the form of money exchange. Soviet enterprises, operating under a system of financial accountability (*khozraschet*), make monetary payments, at centrally established prices, to producing enterprises for the goods they receive from these enterprises. This system was introduced for its administrative contributions rather than for its contributions to economic decision making. Yet it does bring a touch of market atmosphere to the centrally planned, rationing interactions between state enterprises.

In the post-1965 period, there appears to have been some attempts to increase participant discretion in interenterprise transactions. There has been, for example, talk of developing wholesale trade mechanisms (with state established prices) in place of the producers goods

TABLE 1—MATRIX OF HORIZONTAL INTERACTIONS IN THE SOVIET ECONOMY

Producer \ User	State Enterprises	Collectives	Households
State Enterprises	1 A) Rationing B) Red Market (primitive form) C) Gray Market	2 A) Rationing B) Brown Market	3 A) Rationing B) Free Goods & Services C) Red Market D) Pink Market E) Brown Market F) Black Market
Collectives	4 A) Rationing B) White Market C) Brown Market D) Black Market	5 A) White Market (limited)	6 A) White Market
Households	7 A) Rationing B) Red Market C) Gray Market D) Black Market	8 A) Rationing B) White to Light Gray	9 A) White Market B) Gray Market C) Black Market

rationing system. But little has come of this, for many reasons, including shortages of producers' goods in relation to the plan as constructed, and the very difficult problem of devising a set of evaluation and control indicators wherein the various indicators give mutually consistent results leading to the encouragement of desirable behavior and the discouragement of undesirable behavior.

B. *Red Market (primitive form).* Soviet enterprises have the right to sell (redistribute), at established prices, machinery, equipment, and materials that they acquired in the past, but which they no longer need. However, permission from a superior unit is required for the sale of machinery and equipment, and materials in generally short supply.

C. *Gray Market.* Because of supply unreliability and pressure to fulfill output plan targets, Soviet enterprises are staffed with expediters who scour the economy in search of needed supplies, and who employ a variety of monetary and nonmonetary (including goods barter) means to acquire these supplies. This aspect of interenterprise relations is well detailed in the general literature on the Soviet economy; it has often been described as the grease in the horizontal relations that allows the Soviet command economy to function.

Cell #2 (State Enterprises to Collectives)

A *Rationing.* Collectives acquire producers goods, mostly equipment and tools, from state enterprises, by means of centrally planned and rationed allocation through the official material supply system.

B. *Brown Market.* Shortages of spare parts for machinery is an endemic problem in the Soviet economy. Collective farms suffer particularly from this problem. One means of acquiring these needed spare parts is through the bribing of state enterprise officials. This is considered more illegal than the gray market activities in Cell #1 because it involves the unauthorized transfer of state property out of the state sector. On the other hand, it is less illegal than black market activities, because it was not done by officials of the collective farm for personal gain, but for the benefit of the collective. Consequently, in the cases that have been reported, there have been penalties for the collective farm officials involved, but these have been less severe than criminal prosecution. Hence, the designation of brown market.

Cell #3 (State Enterprises to Households)

A. *Rationing*. Except for housing, outright rationing of consumers' goods is not the normal way of distributing consumer goods from state retail stores to consumers. It has been used only in emergency periods, such as the chaotic years of the early 1930's when central planning was introduced, and the period of World War II and its aftermath.

B. *Free Goods and Services*. Health and education services are provided to the Soviet consumer primarily (though not solely) free of charge, but not in unrestricted amounts. Interestingly, however, as Soviet per capita income has increased, there has been a shift of some health and education services from the category of free services to the red market.

C. *Red Market*. This is the basic way in which consumers goods are distributed. Soviet consumers buy the goods and services they desire from state retail stores at state established prices. Consumers' ability to acquire goods is constrained by limited budgets (purchasing power accumulated primarily through sale of labor services), and by limited supplies of goods.

D. *Pink Market*. Individuals are able to resell goods that they purchased previously through a special set of state stores called commission stores. The price to be charged for the good is set by the store official (negotiated with seller), but it cannot be higher than the price of an equivalent good in the state retail store.

E. *Brown Market*. Another endemic problem in the Soviet economy is the (periodic) shortage of individual consumers goods. In response to this, the practice has developed that salespeople in retail stores will keep deficit items "under-the-counter" and will sell them, at raised prices, to certain customers. This practice is very widespread and is punished infrequently and lightly. It can even be argued that the proceeds from such sales are considered by store officials to be part of the expected income of sales personnel.

F. *Black Market*. There are some strictly illegal sales that contribute to the benefit of the retail store as a unit instead of just to an individual employee, but most black market purchases by households fall in Cell #9 (households to households), rather than here.

Cell #4 (Collectives to State Enterprises)

A. *Rationing*. Relevant state enterprises, primarily those in the light and processed food industries, acquire materials (industrial crops, grain, meat, sugar, etc.) from collective farms through the official state material supply system. This includes the above-plan output of these products by the collective farms.

B. *White Market*. Cooperative stores (treated here as state enterprises) buy some food items from collective farms (and individual collective farmers) at freely negotiated prices which they then sell on commission to the public. Such sales comprise about 5 percent of consumer cooperative food sales.

C *Brown Market*. When state enterprises need extra inputs for output plan fulfillment, they resort to unauthorized means of acquisition. These are similar to the gray market activities in Cell #1, but since they usually involve money bribes, they appear to fall in the category of the brown market.

D. *Black Market*. State enterprises at times acquire materials from collective farms through illegal means (bribery) in order to produce goods for sale on the black market, rather than to meet output plan targets.

Cell #5 (Collectives to Collectives)

A. *White Market*. This is not an active cell. The limited transactions that do take place usually involve the redistribution of equipment and tools at freely negotiated prices. It would thus appear that there was no government control, but in reality these transactions are subject to the control of regional Communist Party officials.

Cell #6 (Collectives to Households)

A. *White Market* The collective farm market (similar to farmers' markets in the West) is the prototype of the white market, where the participants are free to set prices. Although there appear to be rules established by superior authorities that prices in the collective farm markets should not be more than twice or three times as high as in the state stores, these rules do not

have the force of law—they can be arbitrarily changed, and are not always enforced. Most of the sellers in the collective farm markets are individual members of the collectives; thus most collective farm market transactions fall in cell #9. But collective farms themselves also sell on these markets, and so some transactions appear here.

Cell #7 (Households to State Enterprises)

A. *Rationing.* Most labor service in the Soviet Union is not allocated through compulsory means by superior authorities, but some are. The part that is includes the fixed period (usually up to three years) of assigned service that must be put in after completion of most professional training. It also includes the armed forces, and in terms of economic significance, especially the construction troops. Finally, it includes the labor camps.

B. *Red Market.* The dominant form of interaction in this cell is the red market. Soviet workers sell their labor services to state enterprises at prices established by the state, which vary by industry and occupation, and within industry and occupation, by workers' skill level and their productivity. Theoretically, a worker is free to sell his labor anywhere, but actually he is constrained by limited availability of housing, job opportunities, and access to requisite training.

C. *Gray Market.* Some elements of gray market are to be seen. In response to the pressure of labor shortage, managers of enterprises often overclassify workers in regard to their skill levels, and thus raise the price they offer for their labor services

D. *Black Market.* Cases have been reported where enterprise officials make illegal payments to workers which are then shared by the officials and workers. Also, fictitious workers ("Dead Souls") are at times included on the work force list and their wages pocketed by enterprise officials.

Cell #8 (Households to Collectives)

A. *Rationing (a form of).* An individual born on a collective farm does not have an internal passport (the announced intention is to change this), and does not have the right to leave the collective farm without permission. Most of the voluntary emigration from collective farms to urban areas is accomplished through arranged nonreturn after compulsory military service.

B. *White to Light Gray.* Collective farms sometimes purchase construction services from small teams of (3–5) workers. These are often students during the summer months, who are aided by party organizations in the acquisition of building materials. There are also other free-lance construction teams, called *shabashniki*, who are frequently Armenian, and who usually manage their own access to building materials.

Cell #9 (Households to Households)

A. *White Market.* This is where most of the transactions in the collective farm market (described in Cell #6, above) take place; these comprise a substantial share of the sales of certain food items in the *USSR*.

Also, if a doctor who has private patients registers and pays taxes on the income derived, his (private) transactions appear here, but it is rare that a private doctor so registers.

B. *Gray Market.* A broad array of goods and services are sold in interhousehold transactions in the semilegal gray market. These include: summer homes *(dachas)*; the services of workmen to build these homes; the services of repairmen to fix consumer durables (including automobiles), plumbing, electricity, etc.; all sorts of private tutoring and lessons; most private doctors; and others.

C. *Black Market.* This is not the place to discuss at length the household to household black market, which forms a major part of the "second economy" existing in the Soviet Union. Suffice it to say, that it is comprised of individuals who steal state property and sell it for private profit, and of individuals who illegally produce consumer goods, usually with state property as inputs and with use of state machinery, and sell these goods for private profit. The latter producing units are often quite large and well organized. Furthermore, it is interesting to note that transactions in this black market in the Soviet Union usually involve legal goods which are in short supply, rather than illegal goods; there is, for example, little trade in narcotics.

II

Having described the terrain of horizontal interactions in the Soviet economy, in what ways may we have contributed to an increased understanding of the plan-market problem? First of all, the complexity of the issue has been underscored. Some of the diversity of "plan" mechanisms has been suggested, but more, the broad scope and wide variety of "market" mechanisms in the Soviet planned economy have been indicated.

This gives rise to the question, why is there such broad use of market-type mechanisms in the Soviet economy? Why are not all the horizontal interactions of the rationing type? A full response to this question is not possible here, but such a response could involve the following elements. The market mechanisms that are observed in the Soviet economy arise for three types of reasons. The first is related to the diversity of needs and behavior of individual units and to the immense amount of information required for centralized planning in a large, developed economy. The Soviet leaders decided that in regard to the distribution of goods among consumers with varying preferences, the computational costs involved in complete rationing vastly outweighed the benefits gained from central control. In the producers' goods markets, the decision was the reverse. Yet current discussions of wholesale trade (red market) in producers' goods reflects the problem of computational costs. Such markets, especially within the consumer sector, are observed in all large, nonprimitive economies. Therefore, they may be referred to as *universal markets*. These not only include the red and pink markets, but also those black markets that deal in goods prohibited in a society for which a demand exists.

A second group of markets has resulted from the Soviet planning system itself and its operational deficiencies, especially the excessive tautness in Soviet plans and the resulting shortages and supply unreliability. Primary among these markets is the gray market in producers' goods. Furthermore, due to shortages caused by planning deficiencies, and due to irresponsible attitudes toward state property (high level of theft of state property), the black market in consumers goods is extensive. This group of markets may be called *Soviet planning markets*.

A third group is comprised of the markets which have appeared in the Soviet Union in connection with the specific historical conditions in the development of an agrarian country with a low standard of living. In this group, which we may call *low level of development markets*, are the white and gray markets for consumers goods, and the black markets involving shortages which are related to low-level of development and which are exacerbated by certain government policies—high rate of investment and artificially low prices on consumers' goods.

The future of these markets depends on many things. But in any case, markets will change as the Soviet economy develops. If we are to assume that in the *USSR*, the standard of living will rise and that the planning system will be improved, especially in regard to the vertical mechanisms and their internal consistency, then we would say that the role of universal markets will increase, and the role of Soviet planning markets and low level of development markets will decrease.

Finally, we have covered only a small part of the plan-market question. In a fuller treatment of the problem, the key issue of balance between plan and market would have to be addressed. This balance is related to the problem of power in the society; it is a function of the degree to which resource allocation is influenced by central planners and controllers, and the degree to which it is affected by discretionary, free decisions of participants in horizontal relations. In terms of the latter, the whole issue of incentives and property rights needs to be examined, in order to analyze the decision-making processes of peripheral units and direct participants in horizontal relations, and the efficiency of these decisions and their compatibility with societal objectives as interpreted by central leaders and planners.

It is clear that we have barely scratched the surface of the plan-market problem in this brief paper. But as the late Chairman Mao once said, "a long journey begins with one small step."

DISCUSSION

PAUL M. SWEEZY, Co-editor *Monthly Review:* I used to think that important principles were involved in the market vs. plan question, but I no longer do. The concept of a pure market economy is of course an extreme and I think not very useful abstraction. Governments and other public or private institutions have always played an important role in the functioning of capitalist economies. I vividly remember Ripley's course on railroads at Harvard many years ago, where we students learned that during the nineteenth century the *U.S* federal government had granted to the railroads a total area bigger than France and Germany combined (I can't vouch for the accuracy of either the fact or my memory), after which I could never believe in the market determination of the economic process in what was then supposed to be one of the two or three freest economies in the world. And those who, like John K. Galbraith and Andreas Papandreou, argue that the present-day *U.S.* economy is largely planned by a few hundred giant corporations can certainly adduce a great deal of evidence to support their view. On the other hand even those countries, like the *USSR*, which officially boast of planned economies make very extensive use of markets to carry out the production and distribution of goods and services. Planning and markets are certainly not mutually exclusive, and on the basis of the historical record it would seem to be more accurate to say that they are complementary rather than antithetical.

Nevertheless, the notion of an opposition between market and plan, while by no means inherent in the concepts as such, *may* refer to a reality. Markets tend not only to reproduce but also to intensify and strengthen the social relations which they embody. They favor the rich over the poor, the skilled over the unskilled, the knowledgeable over the ignorant. If these tendencies agree with the biases of planners, or if planners want at most to moderate excesses in the processes and results of market behavior, there is no conflict. On the other hand, if planners aim to achieve results seriously at odds with the natural tendencies of markets, or rather of a system of markets, then indeed there is real opposition which may result in a struggle with several possible outcomes. But it will not be a struggle of market against plan or vice versa, but rather an essentially political struggle between the beneficiaries of the status quo and those who wish to change the status quo.

I think this gives us a useful framework for interpreting American controversies over planning. Big business, the main supporter and beneficiary of the status quo, is afraid any move toward explicit planning will be controlled by advocates of change and therefore seeks to identify planning with subversion of private enterprise and all manner of related evils. Those on the other side, more specifically the prestigious Initiative Committee for National Economic Planning, are more sophisticated: planning, they say, far from threatening the existing system is urgently needed to strengthen it and improve its functioning in what looks like being a protracted time of troubles. They can and do point to France, the capitalist country that has been a pioneer in developing the technique of "indicative planning" and where, according to a recent Organisation for Economic Co-operation and Development study, the distribution of income is at or close to the top in inequality. What the promoters of planning are saying, in other words, is that as an ideological issue planning is a phoney and that those in business and elsewhere who allow themselves to be taken in by it are depriving themselves of what may be a useful weapon in defense of their own interests. If I understand him correctly, Richard Musgrave would not disagree.

Deborah Milenkovitch's discussion of the Yugoslav case is useful in bringing us up to date on developments in that country. She argues that the market-dominated system adopted in 1965 failed in all its major objectives and has been replaced by a more centralized and controlled though still far from complete system during the 1970's. Limitations of time obviously prevented her from discussing the relative success of this new departure, which is too bad. Instead she speculates on whether the re-

vised system is calculated to promote the goals which Yugoslav theorists have articulated—more autonomy and initiative at the base, greater regional and personal equality in the distribution of income, etc. And her very tentative answer, I take it, is that this is likely to be the case, the reason being that under the new setup the power of the economic elites has been reduced and that of the political organizations increased. The implicit assumption seems to be that the political managers are to the left of the economic managers. This may be so, but I do not think it can be taken for granted. As a working hypothesis I would tend to assume that both sets of managers belong to an emergent ruling class and that the distribution of power between them corresponds to what are perceived to be the overall needs of the system as a whole and hence to the interests of both groups. This at least would help to explain what Milenkovitch finds so puzzling, i.e., "how the allegedly powerful technical, managerial, and financial interests, victors in the struggle for the 1965 reforms and gainers in the redistribution of power after that reform, so readily acceded to a reform that apparently diminished their powers." My guess would be that nothing short of a cultural revolution on the Chinese pattern could now serve to effect a really significant redistribution of power in Yugoslavia

GARY FROMM, National Bureau of Economic Research and Stanford Research Institute: Popular misconceptions be what they may, there is no wholly planned economy in the world today, nor one that wholly relies exclusively on markets. The economies discussed at today's session range along the spectrum from the more highly planned Soviet system to the more free market *U.S.* system, with the Yugoslavs flip-flopping somewhere in between. What is the ideal system? The authors of these papers appear to believe that it will and should lie in the middle ground of mixed systems, partially planned and partially reliant on market decisions.

As Richard Musgrave aptly observes, unfettered and unguided market mechanisms are unlikely to meet requirements, among others, of adequate macro performance, of sufficiently limiting undesired external effects, and of a distribution of income and wealth which accords with society's preferences. At the other pole, a completely planned system might achieve its objectives with respect to income distribution and limitation of undesired external effects, but only at the cost of losses of production stemming from inefficiencies engendered by rigid adherence to detailed specifications of prices and outputs. From the standpoint of maximization of social welfare, elements of both planning and market are needed and, indeed, are utilized in all but the most primitive of economic systems.

The belief held by a sizable segment of the *U.S.* business community that the *U.S.* economy is unplanned largely is a myth. Government intervention and regulation is pervasive and has significant effects on the allocation of resources, the prices of outputs, and the costs of factors of production. Minimum wage laws, occupational safety and health standards, unemployment insurance contributions and payments, equal employment opportunity regulations, and pollution standards, impinge widely on all industries. Transportation, communications, and other public utilities are extensively regulated. Federal lands are leased for mineral, lumber, and oil and gas exploration and production. Agricultural policy is directed toward both price and quantity limitation within lower and upper bounds. Manpower requirements are monitored and, by various means, efforts are made to avert severe long-run surpluses or shortages in selected occupations. Taxes, subsidies, guarantees, tariffs, quotas and other instruments are used to influence supply and demand for broad classes, and for detailed types of goods and services. It is possible to list many other forms of intervention which are designed to, and do affect, resource allocations, in both the short and the long run.

In many cases governmental economic objectives are similar in capitalist and socialist countries, and planning for achievement of those objectives takes place. The mechanisms employed

differ, with capitalist nations resorting more to use of markets and reactions to prices and incentives and socialist nations relying more on direct control or intervention on quantities, with consumer and producer prices administratively set at fixed levels over long intervals. Also, socialist more often than capitalist intervention has been done in depth and detail. However, whatever the mechanisms used, there should be little question that the consequences of all major policy initiatives which might be undertaken should be examined within a comprehensive framework. Otherwise, contradictions are likely to arise and desired degrees of specific or total effects would only occur by happenstance. Stated more technically, sets of simultaneous local maximizations may not be feasible, and even if feasible, are not sufficient to guarantee achievement of a global maximum

Musgrave offers a number of examples of needs for a comprehensive short- and long-run overview and policy cohesion. In order to help achieve such goals simultaneously, he advocates consideration of several alternative organizational formats for an agency which would prepare long-range policy plans. He also raises the question whether a long-term plan should be submitted and enacted on an annual basis or every four years. These institutional characteristics, while important, need not concern us here. They relate to operation of the political process in the United States and are not fundamental to the issue whether planning is necessary to more closely achieve social optima on a wide front of national goals

Musgrave focuses finally on the key point, the need for a coherent set of policies to complement the market. This is more than as he indicates at first a matter of policy semantics. A mixed regime of planning and market, encompassing both freedom of participants (constrained and influenced by government) and forecasting and response to individual and enterprise forces probably would lead to closer adherence to socially desired outcomes than systems that would strive to attain either polar extreme—a pure market or a fully planned, and regimented system.

The United States already has a mixed economic system, but does not operate it well. The evidence for that conclusion should be clear from the performance of the past decade. One of the principal ingredients that appears to be missing in *U.S.* policy formulation is an analytical system that can produce highly detailed, internally consistent, reproducible short- and long-run forecasts of macro- and microeconomic effects of alternative policies. The present apparatus of the Council of Economic Advisers is inadequate for that task, and the Troika has relied too long and heavily on informed judgments and not sufficiently on scientific methodology. The industrial outlook predictions of the Department of Commerce are contradictory across industries and when aggregated, agree only by accident with overall national predictions of the Department's own Bureau of Economic Analysis, or those of the Troika or any other government agency. The black box forecasts of the Federal Reserve, which rarely see the light of day, are so obscure and the prediction process so hidden that we don't even know who the magicians are; under those circumstances, how can we have any confidence that they are analyzing properly the consequences of different monetary policies? Finally, there is the interagency growth analysis group, now almost exclusively Bureau of Labor Statistics personnel, which occasionally has been involved in examining policy issues but could be utilized far more effectively if its efforts were coordinated or combined with those of the Council of Economic Advisers

The impetus for the Humphrey-Hawkins and Humphrey-Javits bills arises from various sources. Clearly, one of them is the inadequacy of the government's economic analysis capabilities. Even short of implementing planning actively, much more can and should be done to strengthen these capabilities so that, at a minimum, economic policy can be formulated on a far sounder basis. This probably entails reorganization of the Council of Economic Advisers, which appears to be long overdue. Musgrave's thoughtful paper tends to support that conclusion.

Toward the other end of the spectrum, Soviet economic analysis capabilities, too, seem deficient and inadequate to the monumental tasks of operating well a planned economy. The paper by Aron Katsenelinboigen and Herbert Levine unfortunately does not address that subject but is confined to categorizing transactions among state enterprises, collectives, and households by their degree of legality in a matrix format The classification is nonparametric and color coded from light to dark to describe the range of markets from legal to illegal. While not new, this is an interesting expository device and should help those unfamiliar with the Soviet economy to understand its operations. In their all too brief final section, the authors give three types of reasons for Soviet partial reliance on markets: 1) The immense amount of information required if all resource transfers were to be centrally directed; 2) the ability to respond to shortages to some degree; and 3) incomplete evolution from an agrarian economy They expect future Soviet dependence on free (universal) market mechanisms to increase, and the role of centrally planned and directed resource allocations to decline. The reasons for this are not given by Katsenelinboigen and Levine This would be rational on economic efficiency grounds, but need not occur if Soviet leaders are adamant about maintaining strict economic and political control. Until now, my impression is that the Soviet economy has not been operated anyplace near its production possibility frontier. There is a little reason to presume that in the near future the Central Committee would approve reforms which would help bring that about, but would tend greatly to diminish its authority.

In the case of Yugoslavia, there has been greater emphasis on efficiency than in other socialist countries and willingness to experiment with greater and lesser degrees of centralized control of the economy. The most recent swing has been toward greater coordination and direction and emphasis on planning. This apparently was done because of dissatisfaction with the income distribution arising from a freer system and to avoid certain gross duplications and inefficiences associated with lack of *ex ante* coordination. How well the new organizational form will operate remains to be seen, since key elements are yet to be defined and the transition to integrated planning is incomplete. The flexibility exhibited by Yugoslavia in operating its system within a "learning by doing" milieu, however, is beneficial and should minimize the risk of large economic and political losses.

All the authors are to be congratulated for fine contributions to the analysis of plan vs. market. Prescriptions for defining an ideal mixed system still remain to be written, and I look forward to reading their further work on this subject.

DISTRIBUTION OF INCOME AND WEALTH

On International Comparisons of Inequality

By GRAHAM PYATT*

At a previous meeting of this Association my colleague, Montek Ahluwalia, presented a paper in which international cross-section data on inequality was correlated with levels of income and other variables which characterize aspects of the development process. This was perhaps the latest contribution in a tradition which goes back at least as far as Simon Kuznets (1955) and can be traced through an impressive literature including Irma Adelman and Cynthia Taft Morris and Felix Paukert. In all cases the authors have been careful to mention that U-shaped (Kuznets) curves and observed correlations should not be seized on as causal relationships And there has been a considerable literature making specific points within this general theme. Particularly worthy of note are contributions pointing out the importance of demographic factors which, operating through income-life cycle relationships, can account for apparently substantial differences in inequality between two societies in which all individuals at birth have in fact precisely the same prospects. Thus there is a growing literature of stylized facts and cautionary tales which has as yet to take us very far in developing the political economy of income distribution in relation to development. In this paper I want to suggest that it may be useful to pose the issues in somewhat different dimensions if an integrated view of growth and equity is eventually to emerge.

First I want to argue that questions of inequality are not restricted to differences between households or individuals but also concern the role and command over resources of other institutions in the economy. Thus distribution of disposable income between institutions, broadly defined, is a prior question to that of distribution within the household sector. This point is made in Section I by considering how income distribution arises in a social accounting context.

Secondly, I want to argue that a prime reason for being interested in inequality is concern for the level of welfare in a society as opposed to the level of income. It then follows that living standards, not income, should be the main concern. This leads to the contention that consumption, not income is the relevant variable for analysis. Arguments in support of this position are given in Section II.

Finally, a brief concluding paragraph makes reference to a number of issues which are opened up by the suggested approach. These can only be given limited treatment here. Meanwhile, however, they provide an agenda for future work on international comparisons of inequality which recognizes that there are dynamic tradeoffs between growth and welfare, thus taking the formalization out of its present comparative static framework.

I. Social Accounting for Income Inequality

While economic theory has real difficulties in attempting to reconcile questions of distribution with those of income levels and their growth, there is no conceptual problem in embracing income distribution questions within a social accounting matrix to show quantitatively "who gets what" within the national product. Table 1 (which reproduces Table 3 of Pyatt and Erik Thorbecke) provides an appropriate format for doing so.

*Senior Adviser, Development Research Center, World Bank The views expressed in this paper are those of the author and should not be attributed to the World Bank or any of its affiliates. Thanks are due to Montek Ahluwalia, Don Keesing, Jacob Meerman and Suresh Tendulkar for comments on issues bearing on the subject matter of this paper.

TABLE 1—TABLEAU FOR THE ANALYSIS OF INCOME DISTRIBUTION

		Institutions Current Accounts			Totals	Production Activities	Rest of the World
		Households	Companies	Government			
Factors of Production	Labor	Allocation of labor income to households			Factorial Distribution of Income	Wages	Net factor incomes received from abroad
	Other	Allocation of surplus of unincorporated enterprises	Allocation of operating surpluses to companies			Operating surpluses	
Totals		Distribution of factorial income over institutions before transfers			*	Value added	Net factor incomes received from abroad
Institutions Current Accounts	Households	Distribution of factorial incomes over institutions *minus* total transfer payments	Profits distributed to domestic households	Actual and imputed current transfers to households	Distribution of Disposable Income		
	Companies			Current transfers to domestic companies			
	Government	Direct taxes on income	Direct taxes on companies plus operating surpluses of state enterprises				
Rest of the World		Net nonfactor incomes paid abroad			Net nonfactor incomes paid abroad		

Note. * = Aggregate income of the domestic factors of production
Source: Pyatt and Thorbecke, Table 3

The table recognizes many production activities, many factors of production, and many types of institutions which include households subdivided into groups by, say, location and/or income level. The table shows how aggregate income of the domestic factors of production translates into a distribution of disposable income across institutions. Taking the asterisk in the table as its center, the north-east quadrant shows how the abscissa derives from domestic product and net factor incomes from abroad. Summing this quadrant across rows yields the factorial distribution of income. In the north-west quadrant these factorial incomes are spread across the institutions which provide factor services, shown here as households, companies and government. A point to note is that the extent to which operating surpluses accrue directly to households depends on the extent to which production is organized by incorporated,

as opposed to unincorporated businesses. Then, through current transfers in the economy (recorded in the south-west quadrant), the aggregate domestic factor income translates into the disposable income of each institution. In addition to government taxation and income subsidies it is important to note that, if all incorporated businesses are state owned, then all their profits are distributed to government. If they are privately owned then only a fraction of these surpluses are distributed, and in this case they accrue to households.

Table 1 makes the point that of the many factors which determine the disposable income of households, important differences between countries will depend on the extent to which productive activities are incorporated, and on the extent to which corporations are state owned. Income inequalities among individuals will therefore depend on the extent of household and unincorporated business activities, since profit from these will inflate the incomes of proprietor households who are most likely richer than their neighbors. For these reasons it is almost certainly the case that international comparisons overstate income inequality in poor countries in comparison with countries with highly developed forms of private corporate capitalism. Comparability would require that retained profits in corporate businesses should be imputed as a part of individual income. Doing so would almost certainly imply that the Kuznets curve phenomenon, whereby inequality tends to diminish as income rises over the upper income range, is overstated and may not in fact exist. It would also imply that inequality would appear to be higher in nonsocialist countries, thus increasing the differential in favor of socialist countries which is otherwise known to maintain.

Putting these speculations aside, an obvious inference from Table 1 is that the proportion of income which accrues to households depends on their role in society vis-à-vis the corporate sector and government.

With respect to government, many of the relevant arguments have been rehearsed by Kuznets (1963). In particular he makes the point that societies differ in the extent to which governments provide services such as health and education free or at subsidized rates. It follows that comparability across countries requires some allowance for this variety. Standard procedure for this is to calculate an imputed income for those who receive such benefits in kind. And within the framework of Table 1 the imputation should take the form of an income transfer from government to households, thus supplementing any actual income transfers such as social security benefits, scholarships, interest payments on national debt, etc. Obviously such calculations call for a high level of econometric skill; and they are likely to have an important impact on perceived inequality since it is rare in any society to find that such benefits are spread evenly over the population. But before such calculations are made, two critical points should be noted.

The first point of caution is that national accounts by convention exclude value added by the capital stock in public administration and services. This omission should be corrected. Corresponding to it, current expenditure on schools is not an adequate measure of the income which should be imputed to those who receive benefits from free public education: some allowance for capital expenditures is also required.

The second cautionary note is that there is no reason to suppose that the social services which governments provide are necessarily set at an optimal level from a welfare point of view. Thus imputed income transfers from government to different households may overstate the benefits which derive from social services Equally there may be an understatement. Hence, there should be no presumption that the distribution after transfers of shares in domestic product at factor cost provides a good measure of inequality from the point of view of living standards.

II. Measures of Living Standards

Many arguments have been presented in the literature as to why income is an inadequate measure of living standards. Yet the obvious

logic of starting with the question of how living standards ought to be measured seems to have been avoided.

Utility derives from consumption, and two individuals have the same living standard if their utility is the same. If we now add the condition that comparisons should assume a common utility function for all individuals, it follows that living standards derive from measures of real consumption, not from measures of real income.

From the above standpoint a number of the objections to income as a measure of living standards fall into place. Thus real consumption comparisons must allow for different prices facing, say, rural and urban households; they must allow for differences in household composition; and they must allow both for nonprice rationing of the type that exists in nonmarket economies, and for availability of nontraded goods which play such a large role in less developed countries.

In considering consumption as an alternative to income in his 1963 paper, Kuznets makes the point that consumption has a direct link to permanent income, which in turn abstracts from questions of income-life cycle relationships. We can now add that since consumption derives from utility maximization over time in the permanent income thesis, it is not only a good proxy for permanent income but may in fact be the preferred variable if both were available.

Of course, measures of permanent income are not readily available, and household surveys are generally thought to give better measures of consumption expenditure than of current income. Hence, both theory and availability of data suggest a clear preference for consumption as a basis for analysis.

If consumption is used as a basis for inequality comparisons, then this in no way obviates the need to allow for imputed benefits, such as free schools, as discussed in the previous section. But what it does do is to abstract questions of current savings from the analysis of inequality. Thus, in the framework of Table 1, incorporation of household production activities would lower perceived household living standards to the extent that accumulation takes place in the form of a retained surplus within companies. This apparent switch is obviously devoid of any real meaning. By focusing on consumption, rather than income, changes in the labels which attach to institutions are no longer allowed to influence the differences in living standards which are identified. Moreover through the link with permanent income, differences in living standards are seen to depend on differences in wealth when this is defined to include human capital. This last point would seem to give a useful emphasis for development of analysis.

III. Welfare and Dynamics

One direction in which the above discussion can be developed is with respect to welfare and the performance of economies over time. Specifically, if living standards derive from a basis in utility theory, then a further step links living standards and welfare Thus, suppose welfare is measured by the average level of utility. Then there must exist a living standard, $\tilde{c}$, which would yield for an individual a level of utility equal to the average. Following Anthony B. Atkinson, $\tilde{c}$ can be referred to as the mean equivalent living standard, and this statistic is an obvious cardinal measure of welfare in a society. Furthermore, if $\bar{c}$ is the average level of living standards, then convex utility functions guarantee that $\tilde{c}$ is always less than $\bar{c}$. Hence we can write an identity

$$\tilde{c} \equiv (1 - I)\,\bar{c} \tag{1}$$

where I is an inequality index which runs from 0 to 1 and is measured by one *minus* the ratio $\tilde{c}/\bar{c}$. Different utility functions will obviously imply different performance measures and associated inequality measures. Conversely, different inequality measures can be interpreted as implying different utility functions.

To go further, it can now be noted that if average living standard is measured by consumption, then it is related to average income through the propensity to save. Denoting the latter by σ, identity (1) can now be written as

$$\tilde{c} \equiv (1 - I)\,(1 - \sigma)\,\bar{y} \tag{2}$$

where $\bar{y}$ is the average level of income.

While there is little prospect of securing agreement on choice of inequality measure (or individual utility function), it may, nevertheless, be of general interest to analyze inequality in the framework of identity (2). This is because the framework gives emphasis to two points which are rarely given adequate stress. These are, firstly, that welfare, inequality and average living standards are not independent issues but, on the contrary, are intimately connected. The second is that questions of welfare or performance should not be treated simply as static problems to be solved by redistributive policies at a point in time. Thus the presence of the savings rate in identity (2) serves to point out that two countries with the same income and welfare measures can be different in important ways: one may have much lower inequality, combined with a higher savings rate. Such a society is clearly in a healthier state than one where savings are low and inequality high.

It is unfortunately not possible to pursue here the question of tradeoffs between I and σ, i.e., between consumption inequality and savings. We know relatively little about these, and not least with respect to forms of political and social institutions. It seems likely, however, that there would be some point in pursuing these questions. Accordingly, the analysis of international comparisons of inequality may go further forward by reference to the identity (2) rather than by looking simply at income inequality and levels. Identity (2) bases welfare on consumption, where it surely belongs. This, then, accommodates differences between countries in the role of the state and private corporate capitalism insofar as these reflect the labeling of activities from an institutional point of view. There may remain, however, important differences with respect to efficiency in both production and distribution. And finally the identity (2) makes it clear that policies for redistribution towards greater equality must be considered in the light of their effects both on current income levels and savings. This is a familiar conclusion to reach. Much of the point of the present paper is to set it in a framework which makes it explicitly. It should be noted, however, that in terms of identity (2) the issue is not income equality versus savings but rather consumption equality. This change in perspective could have important implications for the design of policy.

REFERENCES

Irma Adelman and Cynthia T. Morris, *Economic Growth and Social Equity in Developing Countries*, Stanford 1973.

Montek Ahluwalia, "Income Distribution and Development: Some Stylized Facts," *Amer. Econ. Rev.*, 1976, *66*, 128–35.

Anthony B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, 1970, *2*.

Simon Kuznets, "Economic Growth and Income Inequality," *Amer. Econ. Rev.*, 1955, *65*, 1–28.

———, "Quantitative Aspects of the Economic Growth of Nations: Distribution of Income by Size," *Economic Development and Cultural Change*, 1963, *XI*, No. 2, pt. II.

Felix Paukert, "Income Distribution at Different Levels of Development; a survey of evidence," *International Labour Review*, 1973, *108*.

Graham Pyatt and Erik Thorbecke, *Planning Techniques for a Better Future*, Geneva 1976.

Entrepreneurship, Social Mobility, and Income Redistribution in South India

By E. WAYNE NAFZIGER*

Development economists have been preoccupied with the problem of increasing the size of the *GNP* pie to the relative neglect of its distribution. Despite the recent disenchantment with the viewpoint that all classes share in the benefits of industrial growth, empirical data on the distribution of income, business opportunity, and economic power are in short supply.

Many of the studies on entrepreneurship and economic development, when they are not apologetics for ruthless capitalist exploitation as Paul Baran suggests, extol the achievements of the capitalist entrepreneur without examining class origins and monopoly advantage. Joseph Schumpeter sees entrepreneurs, irrespective of class origin, as heroic figures, with the dream and will to found a private kingdom, to conquer adversity, to achieve success for its own sake, and to experience the joy of creation. For Gustav Papanek (1967), the private entrepreneur, who is frugal, diligent, far-sighted, remarkably able, and willing to take political risks, is in large part responsible for the "success" of Pakistan in achieving rapid industrialization.

Three recent empirical studies have reinforced this conception of the heroic entrepreneur. Orvis Collins and David Moore found that most of the entrepreneurs in medium manufacturing firms in Michigan "clearly moved a long way from the somewhat impoverished economic level of their childhoods." Both Nigerian and Greek industrialists were considered highly upwardly mobile by two other investigators (John Harris and Alec Alexander).

This study offers a perspective on vertical socio-economic mobility, and the differences in economic opportunities between the privileged and underprivileged portions of the population. Unlike previous studies of Indian entrepreneurship, the class (parental economic and occupational status) and caste of entrepreneurs are compared to the general population, and are related to the education, experience, initial capital, and business success (firm's value added or income class) of the entrepreneurs. However, limitations on space rule out consideration of the impact of entrepreneurship on economic growth and issues of government policy, while insufficient data preclude analyses of changes in the distribution of the class origins of entrepreneurs over time, and the indirect effects of entrepreneurial activity on the welfare of lower castes and classes. (See Nafziger 1976.)

Fifty-four entrepreneurs in Visakhapatnam (Vizag), Andhra Pradesh, a rapidly growing municipality of 355,045 people, were interviewed in early 1971. The statistical universe consisted of the 55 private indigenous industrial firms established during 1958–70, and registered with the Industries Department. Each firm has one entrepreneur, the person with the largest capital share.

I. Caste, Family and Social Community

Hindu entrepreneurs are classified according to caste—Brahmins, Kshatriyas, and Vaishyas (twice-born or high castes), Sudras (low castes), and "untouchables" or Harijans (outcastes). Among Vizag Hindus, high castes, with 28 percent of the population and 13 percent of the blue-collar workers, comprised 65 percent of the entrepreneurs. None of the entrepreneurs, but 16 percent of the Hindu blue-collar workers, were Harijans, who were 11 percent of the Hindu population. Table 1 indicates a highly signficant positive relationship between the caste ranking of the Hindu population and representation in entrepreneurial activ-

*Kansas State University. I thank Krishna Akkina, Edgar Bagley, John Nordin, and William Richter

ity. This relationship remains even when the dominant business community, the Vaishyas, is eliminated ($\chi^2 = 10.78$), and even if the analysis is confined to non-Vaishya entrepreneurs born in Andhra Pradesh ($\chi^2 = 8.32$). (Nafziger 1975 indicates the caste composition of entrepreneurs and the population.)

TABLE 1—OBSERVED AND EXPECTED FREQUENCY DISTRIBUTION OF SAMPLE HINDU ENTREPRENEURS BY CASTE RANKING

Caste Ranking	Observed Frequency	Expected Frequency
High castes	28	11.9
Low- and outcastes	15	31.1

Note: Calculated value of $\chi^2 = 30.11$

There is a significant positive relationship between the caste ranking of Vizag Hindus, on the one hand, and education, income, occupational status, and perceived class and status on the other (Kanisetti Ramana). Accordingly, 11 high-caste entrepreneurs indicated a high paternal economic status, 15 a medium status, and none a low status; 2 low-caste entrepreneurs a high status, 8 a medium status, and 4 a low status (calculated value of $\chi^2 = 9.70$). Partly as a consequence, entrepreneurs from twice-born castes ranked substantially higher than Sudras in average initial equity capital (Rs. 191,000 to Rs. 64,000), proportion with a bachelor's degree (57 percent to 20 percent), and the proportion whose major previous occupation involved management responsibility (71 to 33 percent). (The calculated value of χ^2 is significant at the 5 percent level in all instances). The lesser socioeconomic status, initial capital, educational achievement, and entrepreneurial and management experience of Sudra businessmen were significantly correlated with a lower level of median income or value-added (Nafziger 1975).

The Vaishya (mercantile) community, which was 3 percent of the Hindu population of Vizag, comprised 28 percent of the Hindu entrepreneurs. Major Vaishya families, because of their wealth, can provide the training, education, business experience, and capital investment needed for the entrepreneurial development of their sons. If entrepreneurs are classified according to birthplace (whether in-state or out-of-state) and caste, the eight out-of-state Vaishyas ranked highest in paternal economic and occupational status, median education, median prior management experience, median initial capital, median value-added, median income, and the percentage who received the bulk of their initial capital from other family members.

Each of the families of these eight entrepreneurs had at least 7 business units in India (while four families had more than 20). During the colonial period, the families of the eight entered sectors of trade, finance and (in a few cases) manufacturing, in alien linguistic or religious communities, that did not compete with British interests. The capital and business experience amassed became the basis for sizable manufacturing ventures after Independence (1947), when government protected industry and favored indigenous enterprise.

The country-wide network of firms owned by members of a major industrial family maintained a "community of interest," despite recent legislation abolishing the control of several companies by a single managing agency. Major families, because of their resources, knowledge, organizational skill and influence, were more likely to receive licenses for new enterprises or materials, and were better able to take advantage of government schemes encouraging small industry and regional dispersion.

II. Birthplace

Those born outside Andhra Pradesh, who comprised 33 percent of the entrepreneurs but only 6 percent of the population of Vizag, were substantially more successful as businessmen than those born within the state. Outside entrepreneurs were a select group, as those with little wealth and education tended to be thwarted by financial, psychological and linguistic barriers

to interstate migration. In addition, immigrants were more likely to reject local values, obligations and sanctions which impeded rational business practices, and to receive educational and psychological benefits from the challenge of a new environment.

III. Family Status and Resources

The 1969–70 average personal income of the entrepreneur divided by the number dependent upon his income was more than Rs. 2,000, substantially above the all-India average per capita net national product, Rs. 589.3. The income and occupational status of the fathers was high. Accordingly, 20 percent of the fathers were in cultivation, 2 percent in agricultural labor, and 78 percent in nonagriculture, compared to 50 percent, 20 percent, and 30 percent respectively in India's working population in 1951. (Nafziger 1975 The 17 Schumpeterian entrepreneurs or "innovators" had a higher caste and paternal economic status than sample entrepreneurs in general Nafziger 1976)

The economic status of the father was closely related to the entrepreneurial success of the son, in part through the differential access to resources for his investment in education, training, and plant and equipment. Although 63 percent of the male population of Vizag District urban areas had not completed primary school, the median education of the entrepreneurs, who were all male, was some university.

The extended family, because of its age composition and size, frequently mobilized funds that the prospective entrepreneur would not otherwise have available. Sixty-one percent of the entrepreneurs indicated that a part of initial capital was raised from other members of the family. Entrepreneurs with high paternal economic status had an average initial capital of Rs. 319,000, compared to Rs. 66,000 for those with a low and medium economic status. (The calculated value of χ^2 here and above are significant at the 5 percent level.)

IV. Comparative Data

A number of studies of Indian industrialists point to a concentration of entrepreneurial activity among the sons of the members of the large business houses, who represent a small fraction of the population (Nafziger 1971). Although James Berna remarks on the extremely varied backgrounds of sample industrialists in the state of Tamil Nadu, 41 of the 46 Hindu entrepreneurs with some caste designation are from twice-born castes, and the rest are Sudras. A highly disproportionate number of the fathers of manufacturing entrepreneurs in Pakistan, which had a common history with India before 1947, were from traditional business communities while a low percentage of fathers were in wage employment or agriculture. This pattern would suggest, contrary to Papanek's interpretation (1962), that the socioeconomic class status of entrepreneurs was high when compared to the population. Indigenous managers of large public, foreign and private enterprises in India also originate from a highly select portion of the population, as none of the fathers were laborers, only 10 percent were farmers (all of whom were farm operators or owners), and the rest were white-collar workers, government officials, business executives, professional men, and business owners (S. Benjamin Prasad and A. R. Negandhi).

Other evidence suggests that this low degree of class mobility may also be found in much of the rest of the nonsocialist world The proportion of fathers of Filipino manufacturing entrepreneurs from an upper socioeconomic position was 36 times that of the population (John Carroll). Similarly, the fact that in Harris' sample of Nigerian entrepreneurs, 56 percent of their fathers were in the nonagricultural sector compared to 21 percent of the Nigerian male population, and that in Alexander's study of Greek industrialists, 54 percent of their fathers were big merchants, industrialists, professional men or business executives compared to 2 percent of the working population supports the contention that the socio-economic class of the entrepreneurs was far above that of the population at large. (Nafziger, forthcoming 1977. The calculated values of χ^2 here and below are significant

at the 5 percent level.)

Many students of development identify a "rigid" social structure with an underdeveloped economy, and a fluid social structure with a developed economy (Neil Smelser and Symour Lipset, Alexander). But there is no more evidence of upward mobility to business activity in the United States than in India. It is true that two-thirds of the sample entrepreneurs in Michigan studied by Collins and Moore described their early family life as poor or underprivileged. But the only other choices were "affluent" or "well off." Entrepreneurs may also have evaluated parental family income in terms of contemporary standards (the 1960's) and in comparison with their own high level of economic well-being. The vast underrepresentation of Michigan sample fathers in unskilled, semiskilled, clerical, sales and kindred work (24 percent from the sample and 61 percent from the general labor force), together with the disproportional representation of fathers in business, executive, managerial, and official, farm ownership and managerial, and professional work (57 percent compared to 26 percent), points to median incomes substantially above the corresponding population of their period (Nafziger 1975).

Executives in large corporations, who as a group enjoy higher remuneration and lower risk, are likely to originate from an economically more select portion of the population than medium-scale industrial entrepreneurs (Collins and Moore). The *U.S.* business elite in 1952 was comprised largely of "the sons of men of relatively high occupational status, the sons of business and professional men." Among the fathers of this elite, the number of executives or owners of larger businesses was eight times its proportion within the general population, while the number of unskilled or semiskilled laborers was one-sixth its percentage in the population (Lloyd Warner and James Abegglen). There is a rough parallel between the position in the business establishment of white Protestants of western European origin in the United States to that of high-caste Hindus in India, and of Black Americans to Indian untouchables.

Entrepreneurs in Nigeria, Greece, and Michigan underwent, in the aggregate, an intergenerational upward movement (i.e., from father to son) in occupational status and material level of living. However, the fact that the socioeconomic status of the entrepreneurs was higher than that of their fathers does not conflict with the evidence that the paternal economic status of entrepreneurs was higher than that of the population as a whole.

The foregoing discussion should make clear that the heroic model of the entrepreneur presented by previous empirical studies resulted from the types of questions posed. Although the data indicating the upward social and economic mobility of industrialists compared to their fathers are useful, this exclusive focus detracts from the considerable contrast between the socioeconomic class background of entrepreneurs and that of the general population. This suggests that industrial business activity, rather than being a path for substantial upward socioeconomic mobility, is a way of maintaining or defending privileged status, and enhancing or consolidating the high economic position of the family.

REFERENCES

Alec P. Alexander, *Greek Industrialists,* Athens 1964.

Paul A. Baran, *The Political Economy of Growth,* New York 1957.

James J. Berna, *Industrial Entrepreneurship in Madras State,* New York 1960.

John J. Carroll, *The Filipino Manufacturing Entrepreneur,* Ithaca 1965.

Orvis F. Collins and David G. Moore, *The Enterprising Man,* East Lansing 1964.

John R. Harris, "Nigerian Entrepreneurship in Industry," in **P. Kilby, ed.,** *Entrepreneurship and Economic Development,* New York 1971.

E. Wayne Nafziger, *African Capitalism: A*

Case Study in Nigerian Entrepreneurship, Stanford, forthcoming 1977.

———, "Class, Caste and Community of South Indian Industrialists: An Examination of the Horatio Alger Model," *J. Dev. Stud.*, Jan. 1975, *11*, 131–48.

———, "Entreprepreneurship, Social Mobility and Income Redistribution: A Case Study of Industrialists in Visakhapatnam, South India," Honolulu 1976.

———, "Indian Entrepreneurship: A Survey," in Kilby 1971.

Gustav F. Papanek, "The Development of Entrepreneurship," *Amer. Econ. Rev. Proc.*, May 1962, *52*, 45–58.

———, *Pakistan's Development: Social Goals and Private Incentives*, Cambridge, Mass. 1967.

S. Benjamin Prasad and A. R. Negandhi, *Managerialism for Economic Development: Essays on India*, The Hague 1963.

Kanisetti V. Ramana, "Caste and Society in an Andhra Town," Ph.D. diss., University of Illinois 1970.

Joseph A. Schumpeter, *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*, translated by **R. Opie,** New York 1961.

Neil J. Smelser and Seymour M. Lipset, eds., *Social Structure and Mobility in Economic Development*, Chicago 1966.

W. Lloyd Warner and James C. Abegglen, *Occupational Mobility in American Business and Industry*, Minneapolis 1955.

Information Costs, Corporate Hierarchies, and Earnings Inequality

By CHRISTOPHER CLAGUE*

The bulk of economic theory has rested on the assumption of perfect information. Combined with pure competition, its natural ally, perfect information leads to the conclusion that a worker's earnings correspond to his productive contribution. This has been one of the messages of economics which the man in the street finds most difficult to swallow. He is quite skeptical that the corporate accountant making $80,000 a year is "really worth" ten times as much as the ordinary laborer whose annual earnings are only $8,000. One of the advantages of incorporating information costs explicitly into the theory of the earnings distribution is that it provides a coherent explanation for the man-in-the-street's skepticism.

Information costs have played a prominent role in recent theories of labor market phenomena: search theories of unemployment, theories of statistical discrimination, the theory of job market signals (Michael Spence), the lemons principle (George Akerlof 1970), and others. The present paper will be concerned with the distribution of lifetime earnings, and to keep the topic manageable I will ignore racial and sexual discrimination. I will also assume that all employment is in the private sector.

Despite the work on information costs, the notion that earnings correspond to marginal productivity in the long run (apart from discrimination, labor unions, and other imperfections frequently held to be unimportant) is widely accepted in the economics profession. At the least we can say that it appears in textbook discussions of the earnings distribution and in much of the writing of the human capital school. It seems worthwhile, therefore, to spell out the reasons why earnings may not be fully explained by marginal productivity in a world of imperfect information.

The general topic of this paper has been treated by David Starrett and also to some extent by Akerlof (1976). This paper differs from theirs in that our attention will be focused mainly on the role of corporate hierarchies in the earnings distribution. Interest in corporate hierarchies grows naturally out of a concern with information costs. The very existence of the firm is attributable to transactions costs, as Ronald Coase showed in 1937. More recently the existence of hierarchies within firms has been explained with the concepts of bounded rationality, opportunism, information impactedness, and other concepts based on information costs (see Oliver Williamson and references cited therein).

The first point to note about corporate hierarchies is that some of them are successful and others are not. The ingredients of success include the recruiting of high quality personnel, but I would submit that another feature of successful corporations, which has perhaps received too little attention from economists, is an atmosphere within the organization that might be called *esprit de corps*. *Esprit de corps* would not be important were it not for the information costs involved in monitoring employee performance (see the discussion of team spirit in Armen Alchian and Harold Demsetz, pp. 790–791 and the discussion of atmosphere in Williamson, pp. 37–39).

I will now make some empirical assumptions:

1) The creation of *esprit de corps* within a corporation takes time and is chancey.

2) The probability of its creation and its expected intensity are positive functions of the competence and conscientiousness of the em-

*Associate Professor, University of Maryland

ployees.

3) *Esprit de corps* is maintained and strengthened, other things equal, by the successful performance of the corporation.

4) Corporate performance is positively related to *esprit de corps* and to employee competence.

5) Competence and conscientiousness tend to be positively correlated across individuals. In the absence of empirical support for this assumption, I would offer a plausible psychological proposition: competent people find more self-fulfillment in performing up to their capacity than do incompetent people. In addition, competence and conscientiousness may each be positively correlated with social class and hence with each other

The interesting question of how the increased net revenues from improved corporate performance are allocated among employees and stockholders will be addressed below. For the moment let us simply assume that improved performance leads to higher earnings for all employees in the corporation The above assumptions lead to a picture of the private sector of an economy in which some people work in corporations with high *esprit de corps*, others in corporations with less, and some in types of economic activities where close monitoring of employees is possible and *esprit de corps* is unnecessary. Those working in corporations with high *esprit de corps* will tend to be above average in both conscientiousness and competence. The existence of the phenomenon of *esprit de corps* tends to widen earnings differentials.

Another empirical assumption which may be reasonable is that in many circumstances there are positive interactions from putting highly competent people together. This is certainly not always the case There may be little advantage to combining high-quality janitorial services or even high-quality secretarial services with highly competent executives, but it seems plausible that there are positive effects on learning from putting high-calibre management trainees with high-calibre middle managers and the latter with high-calibre executives. Where this positive interaction occurs, successful corporations will take advantage of it and it will further widen earnings differentials.

Those excluded from the productive organizations may be initially only slightly less productive on average than those included. But they are not exposed to opportunities to pick up useful skills and they may pick up unproductive traits from their fellow-workers (Peter Doeringer and Michael Piore). The dual labor market literature gives a variety of reasons to explain why escape from secondary jobs is difficult. The result is that small differences in productive potential can easily become magnified.

In sum, there is a beneficent circle in some corporations leading from high salaries to high-quality people to *esprit de corps* and extra on-the-job learning to high revenues and back again to high salaries. Can the extra earnings be competed away by the creation of rival organizations? The limitations of competition derive, once again, from information costs, at least in part. The successful corporation may have technological secrets or brand loyalty. Recall also that the creation of another organization with *esprit de corps* takes time, is chancey, and requires reasonably high-quality employees.

Now let us return to the question of who receives the benefits of superior corporate performance. With considerable decision-making power resting with the top management, obviously the stockholders need be given only part of the gains. The goals of top management include their own salaries, their working conditions, their power and prestige, and perhaps *esprit de corps* within the organization. In pursuit of these goals, top management will pay rather high salaries all along the line. This facilitates recruiting the kind of people who contribute to *esprit de corps* and it makes life more pleasant for managers themselves (assuming the managers are themselves competent and do not fear competence among their hierarchical inferiors). In addition, top management will be especially generous to itself and to those in the upper portions of the hierarchy. Their salaries may easily be substantially higher than are

required either to induce people within the organization to accept the superior positions or to recruit junior people to the organization. (An interesting discussion of corporate salaries, emphasizing the normative character of past experience and the role of corporate growth in managerial motivation, is contained in Robin Marris, pp. 89–107.)

Let us contrast this picture of earnings inequality with that which would occur under perfect information. In such a world a person's productivity (and earnings) would depend on his ability and on his willingness to acquire training, both in school and on the job. Such a theoretical framework leaves no room for a worker characteristic that is extremely important to employers in the real world—the worker's willingness to perform well even when he is not being closely monitored. In my opinion the role of ability has usually been exaggerated in nonradical[1] economists' discussions of earnings inequality, and the roles of chance and conscientiousness have been underemphasized, both of these being consequences of insufficient attention devoted to information costs.

It is obvious that in the presence of information costs a person's earnings will not normally correspond to the set of productive traits that he brings to the labor market. (I use the phrase "productive traits" in place of "ability." It refers to characteristics of workers when they first enter the labor force.) A person's earnings may be thought of as the sum of a stochastic term and a set of variables representing his productive traits. Let us call this proposition number one. It will probably be readily accepted. It may be regarded as trivial, but I will try to show later that it has some implications which are not always appreciated.

A second proposition, which is more controversial, is that the stochastic term is positively correlated across individuals with a composite index of productive traits. (This index of productive traits is a prediction of lifetime earnings in a world in which the phenomena of esprit de corps and corporate market power were absent.) Another way of stating the proposition is that more productive individuals receive on average a quality bonus. The reasons for thinking this quality bonus exists have been sketched above. People with above-average competence and conscientiousness reinforce each other's productive performance, especially in a world where information costs are of primary importance. In qualification to this argument, it must be mentioned that imperfect information about individuals' performance and potential frequently leads to the payment of identical salaries to people doing very different amounts of work.

In the rest of this paper, the statements made will hold if proposition 1 alone is true, but they will gain added force if proposition 2 is also correct.

Why is it important to acknowledge the role of imperfect information in the theory of the earnings distribution? First, it affects one's interpretation of observed earnings inequality. An individual with very high earnings would be expected to have a positive stochastic term. Moreover, the inequality in earnings exceeds the inequality in the index of productive traits. Recognition of this fact would affect one's predictions about the size of the effects of demographic changes on the earnings distribution (Clague). Second, I think that in the popular mind the principle of payment according to social contribution has some ethical appeal. The attitude of the man in the street toward the justice of redistribution is affected by his perception of the relative importance of chance, productive traits, and the quality bonus in determining earnings inequality. Third, the economist's analysis of the tradeoff between efficiency and equity in designing the income tax is affected by the presence of stochastic elements in the earnings function. This point will require some elaboration.

The recent literature on the optimal income tax (James Mirrlees, Ray Fair, Anthony Atkinson, Edmund Phelps, and Martin Feldstein) assumes that earnings are identical with the value of marginal product. The utility of individuals

[1]Radical economists have made this point. See, for example, Samuel Bowles and Herbert Gintis.

is a function of consumption and leisure (which may be interpreted as the opposite of effort), and social welfare is a function (which may be more or less egalitarian) of individual utilities. The income tax function is selected so as to maximize social welfare. The optimal tax schedule depends on the individual utility function, the egalitarianness of the social welfare function, and the distribution of ability. The resulting tax structure normally provides a minimum income guarantee and is progressive in the average sense; under typical assumptions, however, the average tax burdens are rather low and the marginal rate may fall as income rises.

I would like to suggest that the introduction of a stochastic term into the earnings function would make the optimal tax function more progressive in a marginal sense—that is, it would increase the second derivative of the tax function. Since the subject matter is too complicated for rigorous argument here, Iwill sketch an intuitive justification for the proposition

Suppose the various functions were such that the optimal tax schedule in the absence of a stochastic term in the earnings function, were characterized by a lump sum grant and a constant marginal tax rate.[2] Now consider an increase in the marginal rate for upper income groups, combined with a reduction of the marginal rate to zero for a certain interval of the income distribution, starting at the breakeven point of the negative income tax. This would redistribute income from upper income groups to those just above the breakeven point The change in the marginal rates would provide additional work incentive to the latter group and reduced work incentive to upper income receivers. These effects will be offset to some extent by income effects, but let us suppose for the sake of argument that the income effects are small enough to leave the pattern of incentive changes as described above. The nonstochastic earnings function, by hypothesis, makes the tax change undesirable. The loss in output weighs more heavily in the social welfare function than the redistribution. But if the earnings function were stochastic, the lost output from the upper income groups would be less than their loss in gross income, and this might tip the balance in favor of the redistribution.

This paper has been concerned with the role of information costs and corporate hierarchies in the theory of the earnings distribution. I have suggested that there is a beneficent circle in successful corporations running from high salaries to high-quality people to *esprit de corps* and extra on-the-job learning to high revenues and back again to high salaries. People with less productive traits tend to be left out of this circle and as a result earnings differences are wider than they would be in its absence. To put the point another way, people with more productive traits receive on the average a quality bonus which would not exist in the absence of the phenomena of *esprit de corps* and corporate market power.

[2]The marginal tax rate on the very highest income earned should be zero (Phelps, p 349) but we will disregard this in our example

REFERENCES

George Akerlof, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism, "*Quart. J. Econ.*, Aug. 1970, *84*, 488–500.

———, "The Economics of Caste and of the Rat-Race and Other Woeful Tales," *Quart. J. Econ.*, forthcoming Nov. 1976.

Armen Alchian and Harold Demsetz, "Production, Information Costs, and Economic Organization," *Amer. Econ. Rev.*, Dec. 1972, *62*, 777–95.

Anthony Atkinson, "How Progressive Should Income-Tax Be?", in **Michael Parkin,** ed., *Essays in Modern Economics*, London 1973.

Samuel Bowles and Harold Gintis, *Schooling in Capitalist America*, New York 1976.

Christopher Clague, "The Effects of Marital and Fertility Patterns on the Transmission and Distribution of Wealth," *J. Human Resources*, forthcoming.

Ronald Coase, "The Nature of the Firm," *Economica*, Sept. 1937, *4*, 386–405.

Peter Doeringer and Michael Piore, *Internal Labor Markets and Manpower Analysis*, Lexington, Mass. 1971.

Ray Fair, "The Optimal Distribution of Income," *Quart. J. Econ.*, Nov. 1971, *85*, 551–79.

Martin Feldstein, "On the Optimal Progressivity of the Income Tax," *J. Public Econ.*, Nov. 1973, *2*, 357–76.

Robin Marris, *The Economic Theory of "Managerial" Captialism*, New York 1968.

James Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, 179–208.

Edmund Phelps, "Taxation of Wage Income for Economic Justice," *Quart. J. Econ.*, Aug. 1973, *57*, 331–54.

A. Michael Spence, *Market Signaling*, Cambridge, Mass. 1974.

David Starrett, "Social Institutions, Imperfect Information, and the Distribution of Income," *Quart. J. Econ.*, May 1976, *90*, 261–84.

Oliver Williamson, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York 1975.

ECONOMIC PROBLEMS CONFRONTING HIGHER EDUCATION

Financing Public Higher Education

By WALTER ADAMS*

> The expert may explain why things happen, may even predict what will happen if all the assumptions can be held in place. But for all his science he cannot tell you how to make those decisions which require the weighing of competing claims and aspirations and values
>
> When the whirring of computers and the chatter of committees is done. . do not expect the refinement of specialization to solve the ultimate problems which all experts deposit on the doorstep of wisdom
>
> —Kingman Brewster, Jr

This is not a time of ebullient optimism for higher education. Like the economy, the industry is in recession. Individual institutions are squeezed by escalating costs and lagging revenues; administrators are tormented by "financial stringency"; and the Carnegie Commission speaks darkly of "The New Depression in Higher Education."

The industry, it is widely believed, was overbuilt and oversold during the golden 1960's, only to be overtaken in the 1970's by adverse demographic factors, changing life styles, popular disenchantment, and rampant cost inflation. Roseate expectations, so runs the argument, turned out to be a delusion: a more educated populace did not in fact pave the way to the Great Society; universities did not in fact hold the key to the solution of complex societal problems; and going to college was not a sure ticket to financial success. Therefore, according to the conventional wisdom, the current recession in higher education may well be a secular, not a cyclical phenomenon.

This diagnosis, I submit, is inaccurate and alarmist. Its policy implications for financing public higher education are misleading.

1. *Demand projections for higher education are as naive as exponential population forecasts*. Some examples should make the point:

(a) In 1949, Seymour Harris predicted a serious oversupply of college graduates by the 1960's. He based this prediction on a blithe extrapolation of the experience of the early 1940's

(b) In 1961, the Bureau of Labor Statistics *(BLS)* forecast a 100 percent increase in the demand for scientists and engineers in *R&D* employment by the 1970's. The increase turned out to be 39 percent instead of 100 percent. The *BLS*, it seems, had overestimated probable demand and underestimated the growth in supply.

(c) The late Alan Cartter, perhaps the most expert forecaster of enrollment trends and demand for faculty, came to regard his own predictions as progressively less reliable. His 1964–66 projections were 1-2 percent off; his 1968–69 projections were 4–5 percent off; and his 1970–71 projections were 7–10 percent off.

(d) Contrary to expectations, total enrollments in 1975–76 increased by 9.4 percent over the previous year, the sharpest such increase since the booming 1960's, leaving many institutions poorly prepared to deal with the unanticipated influx of students. And, as if to confound the experts, this year's enrollment is again up by 4.5 percent, i.e., by half a million more students than were counted in last year's surprising total.

The moral is clear: demographic trends, income elasticities, and similar factors, however accurately they may be calculated, are not auto-

*Distinguished University Professor, Michigan State University. I wish to acknowledge the research assistance of Michelle Matel.

matic or mechanistic determinants of college participation rates, and hence of total enrollments. Moreover, the present cannot be safely extrapolated into the future. Forecasters, therefore, must be mindful of the simple truism that tomorrow will not necessarily be the same as today.

2. *The demand for higher education, especially public higher education, has both a "natural" and an "artificial" component.* The first may be subject to autonomous market forces and the iron laws of Friedmanesque economics. The second is a derived demand, dependent on the preference and priority schedules of governmental units, and operating in accordance with Say's Law: the greater the level of state and federal support, the greater the demand for higher education (and, incidentally, the more reassuring the financial health of the higher education industry).

Two observations deserve particular emphasis here:

(a) The "autonomous" demand component, one would suspect, is highly sensitive to prevailing prices, whether they are shadow prices or actual prices. Yet, in spite of the fact that students today are charged almost twice as much for a college education as they were a decade ago—a price increase half again as large as the rise in the general price level—it is remarkable that enrollments have not in fact declined. Of course, it is impossible to determine to what extent the observed price escalation has dampened the potential growth of demand for higher education.

(b) The "artificial" demand component, i.e., state and federal support for higher education, has increased since World War II, at least when measured as a percentage of total personal income. State support has moved sympathetically with rising enrollments, but federal support has responded primarily to national problems and crises. Thus, the aftermath of World War II, the Korean War, and the Vietnam War witnessed a sizable increase in student aid in the form of veterans' benefits. The civil rights revolution brought into being the Basic Opportunity Grants for low-income students. The launching of the Sputnik triggered a spectacular infusion of funds for training, research and construction. But, as the Carnegie Foundation for the Advancement of Teaching points out, "after the initial response, the amount of money that was provided either declined or stabilized as the national concern for such problems either declined or stabilized."

Two examples illustrate this volatility of federal support: in terms of constant dollars, federal expenditures for graduate fellowships increased from roughly $35 million in 1960–61 to a peak of roughly $250 million in 1967–68, only to drop precipitously to $50 million in 1974–75. Similarly, federal construction loans and grants rose from roughly $11 million in 1953–54 to a peak of $1,500 million in 1967–68, only to fall to $210 million in 1974–75.

The implications are clear. In the 1960's the government decided, as a matter of national priority, to win the international competition for the conquest of outer space. It made the necessary commitment of resources to that end. If and when it decides that a similar program of investment in human capital and the mobilization of our research talent is required to cope with today's formidable challenges—energy, environment, health, the cities, coexistence and, indeed, the problem of human survival—we would quickly find that the surplus of trained manpower is in reality a deficit, and that our universities suffer not over- but under-capacity. A new "Manhattan Project" to revitalize the nation's intellectual, scientific, technological and economic capability would quickly convert a demonstrable need into effective demand for the services of higher education. This is a matter of public choice and government priorities, and not of divine will or natural law.

3. *State appropriations for higher education have shown a long-term increase, but appropriations are unduly influenced by the cyclical fluctuations of the economy.*

(a) In some state budgets, according to

Carol VanAlstyne, "The share for higher education has declined from peak levels, but the absolute number of dollars appropriated in the last two years has increased in 49 of 50 states, representing a net increase for all states of 29 percent. Inflation has wiped out about two-thirds of the increase, but even so, real dollar support has increased about 10 percent for all states, with 44 of the 50 states showing real increases."

(b) Nevertheless, the budget policy of many states is perversely tied to the business cycle. Under constitutional constraint to maintain an annually balanced budget, these states tend to increase expenditures in good times (when revenues are relatively ample) and to cut expenditures in bad times (when revenues are sagging). Thus, during the recent recession, according to one estimate, the states reduced annual expenditures by $3.5 billion, cut back or postponed capital outlays by $1 billion, and raised taxes by about $3.5 billion. Such perverse action not only tends to compromise the effectiveness of federal contracyclical policies (as it did in the 1930's), but to wreak havoc with the maintenance of traditional state services. It has a disproportionately adverse impact on higher education which, in periods of recession, must compete for increasingly scarce expenditure dollars with the abnormally growing needs of welfare, relief, and similar social services. It is a competition in which higher education, perceived as a luxury, is not likely to fare very well.

The implication is again clear: the constitutional mandate for annually balanced state budgets should be replaced by a requirement for budgets balanced over the period of a business cycle. Budget stabilization funds should be established which would assure the maintenance of needed services, including higher education, during lean years with revenues accumulated during fat years and saved for precisely that purpose.

4. *The half-hearted commitment of the federal government to higher education, and current methods of financing it by the states, tend to undermine a central function of public higher education, viz. to serve as an instrument of vertical mobility in a democratic society. This is especially so in periods of rapidly rising prices for tuition and auxiliary services.*

In contrast to Europe, American state universities and particularly the land-grant colleges were founded on a radical concept. They were a challenge to the old order and declared for a new kind of education—education for the people as a whole, not the privileged classes alone. They were to provide education for the sons and daughters of the nation's yeomanry whose access to the elite private institutions was blocked by financial entry barriers. These institutions, in the words of Joseph R. Williams, the first president of Michigan State University, were to be "good enough for the proudest and cheap enough for the poorest." They were to extend the possibility of higher education to the economically underprivileged, while providing training in an ever-widening span of fields and bringing the fruits of university research to the nation's fields and factories. In this way, the public institutions were to serve society and to make vertical mobility more than just an American dream.

Although the state universities, land grant colleges, and lately the community colleges have undoubtedly contributed to this goal, it is nevertheless true that the recent escalation in tuition levels has detracted from their achievements. So has the regressive financing of public higher education. As Theodore W. Schultz points out, "the financing of higher education is in general quite regressive . . . because it adds to the value of the human capital of those who attend college relative to those who do not go to college, because it increases the lifetime earnings of college graduates in part at the expense of others, and closely related, because higher education provides educational services predominantly for students from middle and upper income families and a part of the cost of these educational services is paid for by taxes on poor families." In short, says Schultz, "the financing is such that substantial amounts of valuable assets are being transferred by society to a particular intellectually elite set of individuals." (In his study of the "grants economy," the segment of the economy which deals with one-way transfers of exchangeables, Kenneth Boulding comes to the same conclusion. Subsidies to the

state universities, he says, "aid the rich and the middle class.")

TABLE 1—INCIDENCE OF REVENUES AND HIGHER EDUCATION EXPENDITURES BY INCOME GROUPS IN MICHIGAN, 1970

Income Brackets (000's of Dollars)	State and Local Taxes as a Percentage of Income	Distribution of State and Local Tax Burden (Percent)	Distribution of "Specific Goods" Benefits of Higher Education Expenditures (Percent)[a]
Under 1	49.36	1.07	1.07
1–2	19.14	2.00	1.22
2–3	15.47	2.05	.92
3–4	13.82	2.21	2.98
4–5	12.90	2.44	2.71
5–6	12.16	2.90	4.88
6–7	12.03	3.65	5.24
7–8	11.12	4.60	6.25
8–9	10.63	5.50	6.99
9–10	10.27	5.78	6.78
10–12	10.13	12.60	14.79
12–15	9.41	16.50	17.03
15–25	8.56	26.09	23.29
Over 25	7.75	12.60	5.84
		99.99	99.99

Source: Douglas B. Roberts, *Incidence of State and Local Taxes in Michigan* (Michigan State University, unpublished Ph.D. dissertation, 1975), and Donald M. Peppard, Jr., *Public Expenditures Incidence in Michigan, 1970* (Michigan State University, unpublished Ph.D. dissertation, 1975).

[a]"Specific goods" benefits are defined as privately captured benefits accruing to students and their families.

Table 1 illustrates the point with respect to the State of Michigan. It shows that, except for the top income bracket, the distribution of higher education benefits is even more regressive than the regressive state and local tax system. Consistent with this finding, the National Commission on Postsecondary Education reports that, for the nation as a whole, the college "participation rates for 18–24 year olds whose family income is $10,000 or more is twice the rate of those from families with annual incomes of less than $10,000. The total number of students from families with incomes under $10,000 would have to increase 50 percent beyond the 1972 level to reach the same participation rate as the entire traditional college-age population." The Commission also reports that, in 1972, 64 percent of all undergraduates from the $25,000-plus family income bracket, and 75 percent of all students in the $15-25,000 family income bracket, were enrolled in *public* institutions. W. Lee Hansen and Burton Weisbrod reach the same conclusion. "Public higher education subsidies," they say, "go overwhelmingly to young people from middle and upper income families." In short, the available evidence leaves little doubt about the unequal distribution of benefits among income groups from public expenditures on higher education.

What, then, are the policy implications? Assuming that it is our goal to promote vertical mobility through greater access to higher education, as well as to promote greater distributional equity in revenue/expenditure patterns, the following measures would constitute steps in the right direction: (a) correction of the monotonically and significantly regressive state and local tax systems—a measure which can, of course, be justified on broader social welfare grounds; (b) the adoption at public colleges and universities of tuition schedules, calibrated progressively to family income; (c) added inducements for qualified low-income students to attend college, including the offer of "full ride" privileges normally reserved to talented athletes; (d) creation of an educational bank, primarily to relieve financial pressures on the lower middle class, which would lend any qualified student the full cost of his/her education under conditions of lifetime repayment through, for example, a surcharge on the income tax. The adoption of some such a loan proposal, as Boulding suggests, would mean "treating education as a rather peculiar kind of investment, which it is, and hence getting it out from under the pure grants economy." It would also correct some of the distributional inequities inherent in the current method of financing higher education.

Conclusion. In sum, I submit that the reports of the imminent secular decline in higher education are somewhat premature. If that decline occurs, it will not be the result of natural, inevitable, or inexorable forces. Rather, it will reflect a lack of vision, an absence of leadership, and a failure of will. The fault will lie not in our stars, but in our values and priorities.

The Benefits and Burdens of Federal Financial Assistance to Higher Education

By EARL F. CHEIT*

For more than a century, the federal government has been an important investor in higher education. In the dismal Civil War year 1862, the federal government, through the Land Grant Act, made available to the states over 11 million acres of land "to promote the liberal and practical education of the industrial classes in the several pursuits and professions of life." It was an act that moved Andrew D. White, first president of Cornell University, to say, "Since the Romans quietly bought and sold lands on which the Carthaginians were encamped in the neighborhood of the Eternal City, there has been no more noble exhibition of faith in the destiny of a republic." It was only the beginning of a long-term, incredibly successful national investment

Gradually, as the nation's need for the products of that investment grew, the federal government became a consumer of higher education as well It bought research, specialized services and, through student aid, instruction But unlike governments in some countries and unlike other supporters of education in this country—the states, students, and private donors—the federal government showed little inclination to control the colleges and universities its investments and purchases helped to fund. Even in World War II, when the federal government urgently needed educational institutions for research and instructional services, it extended its role as consumer, but except for the service academies, it avoided the role of controller.

Since World War II, partly in response to the urging of college and university officials, the federal government greatly expanded its role as consumer and investor by taking financial responsibility for: 1) the promotion of equal opportunity; and 2) the support of research and graduate instruction. Large amounts of money have been involved. Federal outlays relating to higher education rose from $500 million in 1951–52 to an estimated $10.1 billion in 1976, from .15 percent of the nation's *GNP* to .62 percent

From time to time the leaders of higher education became worried that federal funds would bring federal control. When the issue arose, they spoke up. They wanted indirect costs recovered for research to come in flexible form so that they could allocate it. Some opposed institutional grants in the 1972 Amendments to the Higher Education Act because in this form, federal money might weaken the autonomy of colleges For similar reasons, many of them worked to block the recommendation of the National Commission on Financing Post-secondary Education that the federal government gather unit cost data on colleges and universities

When it came to getting federal money, John Gardner once recalled from his service as Secretary of the Department of Health, Education and Welfare *(HEW)*, that college and university presidents were particular. The method they most preferred was what he called "leave it on the stump."

There is moss on the stump today. In the last ten years, the process of getting and using federal funds has become increasingly burdensome. As a result, an important, agreeable patronage has degenerated into an adversary relationship. A few institutions have publically refused to bend to the federal will—Brigham Young University, Hillsdale College, and Wabash College, for example In a full-page advertisement, the presidents of four universities in Washington, D.C.—American, Catholic,

*Dean, School of Business Administration, University of California, Berkeley

George Washington, and Georgetown—declared the need for independence from increasing federal control. The Presidents of both Harvard and Yale warned their alumni and friends that in its evolving form, federal patronage poses a serious threat to higher education, one of the most serious threats of the next several decades.

Most leaders of higher education have protested among themselves, though not publicly. They are concerned about the problem, but seem to be in conflict about joining a public fight over federal control. Why in conflict about such an important principle? Because the issue has evolved from objectives about which there is little or no disagreement; because the benefits of federal financial assistance obviously exceed its burdens; and because these same college officials are still seeking more federal funds

Four developments shape the new situation:

1) *The federal government has extended its role to controller*. Federal money has always carried with it regulation to assure that program purposes were followed, and that money was legally used. As a basis for regulation, both concepts have been extended. The concept of legal use has been extended to include nondiscrimination and affirmative action. The concept of program purpose has been enlarged to mean that federal financial assistance to one program purpose subjects any other purpose to regulation.

The process began in the mid-1960's. Federal authority was extended through the enforcement language of important civil rights and equal opportunity legislation. Thus the Civil Rights Act (1964) provided:

> Each Federal department and agency which is empowered to extend Federal financial assistance to any program or activity . . . is authorized and directed to effectuate the provisions of Section 601 [on nondiscrimination] with respect to such program or activity by issuing rules, regulations, or orders

Executive Order 11246 (1965), moving beyond the Civil Rights Act, extended federal authority beyond nondiscrimination in the conditions of work on a specific program or activity. It prohibited discrimination on the basis of race, color, religion or national origin on all work being done by a contracting agency. In addition, it required a program of affirmative action. Nondiscrimination on the basis of sex was added by Executive Order in 1967. Complaints filed under this new authority provoked the first enforcement of the affirmative action requirement.

As the regulations and court decisions began to take effect in the early 1970's, the colleges and universities were being moved to another stage of regulation by passage of the Buckley Amendment (1974). Its nominal effect was to authorize student access to educational records. Its main significance is that it further establishes the federal role of controller, explicitly conditioning federal funding on actions not connected with the program being funded.

In the meantime, the colleges have been brought under an increasing number of other socially mandated programs relating to conditions of work and internal operations, such as the Occupational Safety and Health Act, and the Employment Retirement Income Security Act.

The process is moving toward a new stage. The psychological set of enforcement and control is extending through the whole range of federal relationships with colleges and universities. Now even without a change in statutory requirements, such as in administration of research grants by *HEW*, there is a new posture of regulation and control.

2) *Bureaucracy is the mechanism of control and its intrusion into college and university life has been disruptive and expensive*. Justice Brandeis' admonition that "the greatest dangers to liberty lurk in insidious encroachment of men of zeal, well meaning but without understanding" is no longer just a favorite of the business press. Colleges and universities are not strangers to bureaucracy but the impact of the new laws created a large campus corps of instant Jeffersonians. Critics charge that the new regulators are uninformed, cumbersome,

and hostile. They require the gathering of useless data; they cause long inexplicable delays; they play "cat and mouse" games over enforcement; they conduct endless reviews. Sometimes, after periods of indecision the decisions they do make are uninformed about the educational process. It has apparently come as news to some GS 12's that a library is needed for research.

A second concern about bureaucracy is that, checked neither by market forces nor by policy, it moves on inexorably The process of control is extending. Principal investigators, for example, now face intensive "disallowance audits" of their research administration.

Bureaucracy can also move vindictively. A *Change* magazine survey produced confidential responses from officials whose institutions had suffered under an arbitrary order but did not publically complain for fear that a new one would be issued.

Finally, it is expensive At U S Senate Sub-Committee hearings on the issue, one college president was quoted as saying: "Everytime I have to hire a lawyer, I have to turn down an appointment of one Associate Professor." No one knows the exact cost. A study by the American Council on Education (*ACE*) found that a sample of institutions spent one to four percent of their operating budgets on socially mandated programs and that these costs are rising, but the study included programs beyond those involving the regulation issue.

In addition to the expense of dealing with the rules, there are costs in creating the kind of internal organization required by the external bureaucracy. This added cost is changing the environment of colleges and universities, and not for the better.

3) *Federal control is growing, but federal financial assistance is declining, both in relative importance to institutions and in total real dollars appropriated.* In recent years, as Figure 1 shows, the portion of federal funds represented in institutional income has dropped sharply from 23.1 percent in 1963–64, to 15.7 percent in 1973–74. Ironically, the worry about weakened institutional autonomy is proving true without the balm of institutional grants and with relatively less federal money in the school budget.

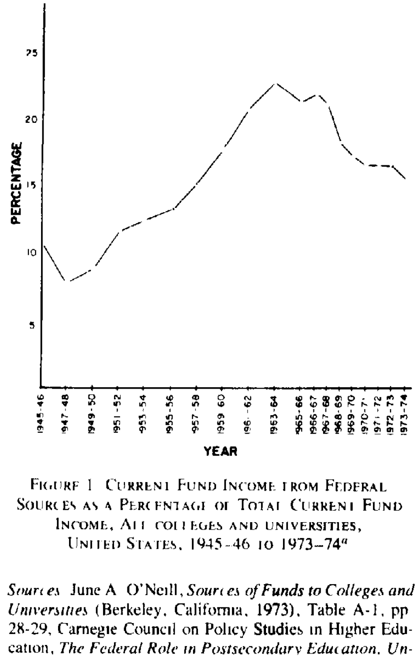


FIGURE 1 CURRENT FUND INCOME FROM FEDERAL SOURCES AS A PERCENTAGE OF TOTAL CURRENT FUND INCOME, ALL COLLEGES AND UNIVERSITIES, UNITED STATES, 1945-46 TO 1973–74[a]

Sources June A O'Neill, *Sources of Funds to Colleges and Universities* (Berkeley, California, 1973), Table A-1, pp 28-29, Carnegie Council on Policy Studies in Higher Education, *The Federal Role in Postsecondary Education, Unfinished Business 1975-1980* (San Francisco, 1975), Table 2, p 71; and estimates based on data from the U S National Center for Education Statistics

[a]The numerator is an estimate of current fund income from the federal government of all institutions of higher education, continental United States through 1951-52, aggregate United States 1953-54 through 1973-74, veteran tuition and fees are not included The denominator is the total current fund income of all institutions

To be sure, not all federal funds go directly to institutions, and some new federal programs are providing funds to the states and to individuals. But although the total dollars appropriated for these and other higher education activities have risen, federal appropriations in constant dollars actually declined about one-half billion between 1973–74 and 1974–75.

4) *College and university officials are most concerned about what will happen if present trends continue*. The fear is that the burdens of federal financial assistance might begin to match their benefits. "The experience of recent years gives fair warning," Kingman Brewster advised Yale alumni, "that reliance upon government support for *any* university activity may subject the *entire* university to conditions and requirements which can undermine the capacity of faculty and trustees to chart the institution's destiny." These fears are not unfounded. In the hope of turning out more family physicians, Congress seriously considered using the leverage of federal money to change the curriculum of all *U.S.* medical schools

Reports that the colleges are fighting federal control stir bemused sympathy in businessmen They have been making a similar case for a long time, often in response to a professor arguing for more regulation of business. Derek Bok's warning to Harvard alumni prompted a physician to write in the *Washington Post*, "The professors have joined the rest of us storeowners, manufacturers, farmers, and doctors who wrestle with federal regulations. Welcome to the party, boys and girls, and shut the door tightly behind you "

Whether the indictment of federal control over higher education fits or is overdrawn is itself a subject of controversy Since most of the evidence comes from examples cited by the colleges themselves, it would seem sensible to keep in mind the old Yiddish maxim—"For example is not proof." Still the accumulated effect of these examples has been persuasive enough to generate serious efforts to reduce the burden of the federal role.

One approach would be to make specific regulatory processes work better, to rationalize the rules, make them clearer, less arbitrary. The Carnegie Council's work on affirmative action makes recommendations of this sort

Another approach, a plan to avoid arbitrary regulations in the future by soliciting greater participation in their formulation, was announced by *HEW* in July 1976.

In the meantime, because the burdens of federal financial assistance are growing, most attention centers on obtaining additional funds to reimburse the costs of carrying out federal programs. The problem has been considered by the U.S. Senate Sub-Committee on Education, but it has not recommended a solution.

Although the cost issue is highly complex, two points are clear: dollar costs are not being fully reimbursed and key nondollar costs cannot be. The costs of research and of training and public service programs are substantially but not totally reimbursed. The outlays associated with student aid programs are, at best, only partially reimbursed. Most of the costs of leverage—the costs, to paraphrase Brewster, of federal police power following the dollar—are not reimbursed.

Money, however, is not the only cost Increased bureaucracy required because of government regulation is another cost, as is the restricted autonomy that comes with federal leverage, perhaps the highest price being paid for federal assistance

Another approach to the problem holds longer term promise. The *HEW* Office of Regulatory Review is considering the creation of a Higher Education Task Force. It could make a helpful review of the problem of federal intervention in college affairs.

Reviews of federal regulation usually result in a triumphant announcement that the number of required forms will be cut, although those that remain will have to be a bit more complex. Such a result is more a measure of the task than of the aspiration A study at the University of California found that 229 "unique reports" are regularly sent to 32 federal agencies. Form cutting alone would help. But a more serious review is required.

The states have an important stake in such a review—they plan around federal initiatives, and would like more stability in the federal role. They and the colleges and universities are subject to federal initiatives, even those not intended to affect higher education. An example is the impact on colleges and universities of social security legislation, analyzed in the *ACE* study referred to earlier

Behind the federal intervention problem is a complex organizational problem. No single entity called the federal government deals with colleges and universities. Most of the funds come from several key agencies, but according to the only recent comprehensive count, 60 different agencies administer 375 separate federal programs whose funds support some activity in higher education. Most of these agencies deal with colleges the same way individual private donors do, one on one. Through these agencies the new posture of control is emerging without a plan, or, in truth, much thought at all. The Buckley amendment was passed without findings, hearings, or even a committee report. Ideally, what is needed is a coherent restatement of the relationship of higher education to the federal government. Neither college officials nor government officials are happy with their present relationship. It is probably too much to hope that a new theory can be formulated, or government structure revised, to everyone's satisfaction. But several important things can be done.

Attitudes can be changed. An adversary relationship between federal agencies and higher education need not exist. Enforcement is a highly specialized function. It is unnecessary and undesirable that a psychological set of enforcement and control influence many federal agencies. A change in official attitudes can remove much of the hostility toward the campus generated in Washington during the Vietnam War and now becoming institutionalized in the permanent bureaucracy.

Regulations can be cut back to those clearly required by law. It is repeatedly charged that actual regulation exceeds that required by law. It would appear that some significant pruning could be done.

The accountability movement can itself be made accountable. Information is not a free good. Not even an inexpensive one. Inevitably, a serious review of regulation must find ways to limit demands for information, make regulation simpler, less arbitrary and costly.

Reimbursement can be made. Dollar outlay costs are a serious issue, especially as their amounts are considerable. The campuses have now had time to assess the real cost impacts of federal programs and can now show what these are. The next step is appropriate reimbursement.

Supply policies can be adopted. Experience with affirmative action shows that restructuring demand raises both consciousness and the advertising revenues of journals, but it does little to raise supply. New policies directed at increasing supply such as those recommended by the Carnegie Council and others, should be implemented.

The case for special regulatory treatment of higher education can be made. Most academics strongly believe that where government intervention is concerned, the market for ideas is fundamentally different from the market for goods; however they have not been energetic in arguing that case. Now it must be considered if the problem of federal regulation is to be dealt with.

Important as it is to work for these six objectives, I believe it is equally important to understand that the mutually beneficial relationship of federal government to colleges and universities that has existed for almost a century goes beyond them. That relationship developed because government's actions, including those in which the government acted as consumer, were based on the theory that higher education expenditures were fundamentally an investment. The federal policies that shaped higher education were guided by a commitment to expand access and a belief that the nation's welfare depends importantly on higher education, not only because expanded enrollment serves social justice and provides an educated citizenry, but also because a capacity for advanced study and research helps the nation meet important needs, some currently identified and others yet unknown.

To benefit most from what colleges and universities can do, government needs more than a restrained, generally agreeable philosophy of control. It needs future policy derived from its old and successful theory of investment.

REFERENCES

Derek Bok, *Harvard University: The President's Report 1974–75,* Cambridge, Mass. 1976.

Kingman Brewster, Jr., *Yale University: The Report of the President, 1974–75,* New Haven 1976.

Earl F. Cheit, "What Price Accountability?" *Change, 7,* Nov. 1975, 30–34, 60–61.

Pamela Christoffel, *A Compilation of Federal Programs Financing Postsecondary Education.* National Commission on the Financing of Postsecondary Education, Washington 1973.

Ronald H. Coase, "The Market for Goods and the Market for Ideas," *Amer. Econ. Rev. Proc.,* May 1974, *64,* 384–91.

Alfred B. Fitt, "The Buckley Amendment: Understanding It, Living with It," *The College Board Review,* Summer 1975, 2–5.

Robert W. Hartman, "The Rationale for Federal Support for Higher Education," General Series Reprint 283, The Brookings Institution, Washington 1974.

Carol Van Alstyne and Sharon L. Coldren, *The Costs of Implementing Federally Mandated Social Programs at Colleges and Universities,* Policy Analysis Service, American Council on Education, June,1976.

Carnegie Council on Policy Studies in Higher Education, *The Federal Role in Postsecondary Education: Unfinished Business 1975–1980,* San Francisco 1975.

———, *Making Affirmative Action Work,* San Francisco 1975.

The Chronicle of Higher Education, "A 1976 Declaration of Independence by the Presidents of The American University, The Catholic University of America, The George Washington University, and Georgetown University," April 19, 1976, 5.

New York Times, "HEW Opening Rule-Making Procedure to Public From Start," July 25, 1976, 18.

U.S. Senate, Committee on Labor and Public Welfare, Subcommittee on Education, *Hearings on Higher Education Legislation,* 1975; oversight hearings on implementation of Federal higher education programs and policy under the 1972 amendments to the Higher Education Act of 1965, June and July 1975, parts 1 and 2.

University of California, Management Information Systems Task Force, *Administration Information Systems in the University of California,* April 1976, Appendix F, Table F-1.

The Washington Post, "Derek Bok's 'Indictment': A Landmark in Academic Policy," June 1, 1976, A–19.

Economic Problems Confronting Higher Education: An Institutional Perspective

By WILLIAM G. BOWEN*

When Robert Strotz invited me to participate in this session, he explained—gently but unmistakably—that I was not expected to produce propositions either original or profound, that being the province of all of you who have maintained your loyalty to the discipline and have not had your gray cells damaged by excessive contact with administrative responsibilites My function, rather, is to share with you a few observations on the economic problems of higher education as seen from an institutional perspective, or, more precisely, from the inevitably particularistic perspective of the president of one university.

The immediate temptation is to assume the role of the coach and produce the crying towel There is, after all, much to cry about, especially when one sees, day to day, the effects on programs and people of resources insufficient to meet many pressing needs. But I would like to try to do a little more than that I would like to think with you about some of our shared problems with the further objective of suggesting a few ways in which simple concepts which derive from our discipline can be helpful in improving understanding.

I. Quality and Productivity

One simple thing economists learn early is that there are many ways in which institutions of all kinds adjust to financial pressures Thus, it is not surprising that in spite of the widely publicized financial difficulties that have affected higher education in recent years there have been very few instances in which institutions have closed down. The instinct to survive is a powerful one, in organizations no less than in individuals, animals, and organisms of all kinds. Indeed, one danger is that too many institutions—and too many programs within institutions—may survive when the price of survival is a level or scale of operation that fails to meet reasonable criteria of educational effectiveness. A second danger is that simply because institutions survive, people may think they are all right.

Of course there are great differences among institutions, within both the public and private sectors, in the resources available to them and in the obligations they are expected to meet. In general, however, my impression is that the great danger is not so much institutional extinction, or even that there will be a sudden, dramatic downward shift from one level of quality to another. The greater danger, I believe, is that there will be a slow, unspectacular, but cumulative decline in what it is possible to achieve—and then, as a next step in the process, in what one *tries* to achieve Gradual changes of this sort are, in their nature, impossible to measure with any precision, and they may not even be noticeable to quite experienced observers until some considerable time after they have occurred.

In assessing what is happening to higher education, economists can be particularly helpful in warning against measures of output (and of "productivity") that ignore important qualitative considerations. Productivity in education and research is notoriously difficult to estimate because quality does matter, whether one is talking about education at undergraduate or graduate levels, or about scholarship and research in fields as diverse as art history and biochemistry. Also, the presence of joint costs, of substantial fixed costs which are hard to allocate among programs and purposes, and of significant economies of scale all complicate the collection and interpretation of data.

*Princeton University.

This is certainly no plea for the avoidance of quantitative analysis. Hard decisions—often concerning things not to be done that institutions could do well—have to be made, and they can be informed by data of many kinds. But we cannot, responsibly and in good conscience, follow a simple "cost per student" test in deciding where to make savings. There are departments and programs at my own university, for example, which are of outstanding quality, which serve national and even international needs, and yet which do not, and will never, attract enough students to come close to meeting any norm defined in terms of the ratio of students per faculty member or cost per student. One example is Astrophysical Sciences, in which Princeton has a preeminent reputation for graduate instruction and research—and five undergraduate majors. An extremely small undergraduate program is inevitable, since even those undergraduates who think they may want to do graduate work in astrophysics are often encouraged to major in mathematics or physics as undergraduates.

In the few minutes I have today, I shall not describe again figures which give at least a rough indication of the reduction in the real value of resources devoted to both education and research at many institutions these last few years. Suffice it to say that, while the specific figures vary depending on the activity and institution and on the price deflator chosen, the general pattern is hard to mistake: Dollar expenditures often have not risen as rapidly as would have been necessary simply to maintain the real value of expenditures per student or per "unit of research" (somehow measured). I do want to say just a word, however, about the interpretation to be put on such figures and about implications of a continuation of recent trends for the long-term quality of higher education.

Those of us who make the case for higher education are well advised, I think, to acknowledge that the great financial pressures of the last decade have led to some helpful increases in the efficiency with which we use our resources. Following the "boom" period of the late 1950's and early to mid-1960's, some tightening of belts in higher education was no doubt salutary, and I would not want to claim that every dollar pared from institutional budgets has represented an equivalent cut in educational muscle. Thus, figures showing declines in the real value of expenditures per student or per research worker may well overstate the actual drop in what it has been possible to accomplish.

At the same time, it would be just as serious an error—probably more serious—to assume that because it has been possible in many instances to adjust to budget reductions these last ten years or so without unacceptable sacrifices of quality, it will be possible to do so indefinitely. Many savings achieved have been of the once-and-for-all variety and cannot be duplicated in each new period. There is a limit to the number of additional windows that can be found each year to be left unwashed. More generally, it is one thing to accept a reasonable period of consolidation or retrenchment which has an end to it and quite another to be unable over some longer period to respond positively to the most meritorious new ideas and new initiatives. There is a serious risk involved in trying to live too long off of accumulated intellectual as well as physical capital. It is hard—and wrong—to take satisfaction from the achievement of a financial equilibrium which makes little allowance, if any, for doing more than preserving the status quo.[1]

There is a related danger. Economists often talk about time horizons, and one of my worries is that too many of us in colleges and universities, as well as outside them, will seek to achieve a temporary financial equilibrium by sacrificing the future. There are of course many ways this can be done, ranging from allowing the physical plant to deteriorate, to spending endowment at an unsustainable rate, to not build-

[1]Moreover, as a number of economists have noted, simply maintaining the status quo is likely to require, over the long run, a rate of increase in expenditures greater than the overall rate of increase in costs for the economy as a whole. The reason, of course, is the labor intensive nature of higher education, and the greater difficulty in achieving real gains in productivity in this industry than in many others.

ing libraries with an eye to the long-term quality of the collections, to not being tough enough now with respect to tenure decisions to insure that there will be at least a minimal number of permanent positions open to the best people in five, ten, and fifteen years' time. Economists should understand these tradeoffs better than many other faculty members, and it is my hope that members of our profession will help to promote a reasonable degree of internal discipline at the same time that all of us struggle to find the new resources that are essential if the quality and vitality of the enterprise are not to be damaged seriously over the long run

II. Equality of Opportunity and Access

I want to turn now to a second set of concerns, having to do with equality of opportunity and access. One of the peculiarities of higher education as an industry is that we care greatly about the range of people able to benefit from the educational opportunities we offer, and not just about their number It is for this reason that so much money, from public and private sources, is devoted to student aid. Yet, programs of student aid notwithstanding, there is serious concern these days that the steep increases in tuition and other charges which colleges and universities have had to impose, and seem likely to continue to have to impose, will deprive qualified students of access to some colleges and universities. This is a vast subject, which I believe deserves more sustained attention from economists than it has received. In the limited time available today, let me offer only four observations.

First, since almost all alumni, admissions officers, high school counsellors, parents, and journalists know of particular cases in which high charges are said to have discouraged someone from attending a college or university (or from pursuing any form of higher education), anecdotal information abounds, and one of the responsibilites of economists, I believe, is to discourage the easy tendency to base sweeping conclusions on such grounds. Higher tuition charges are bound, *ceteris paribus*, to have some deterring effect, and the important questions have to do with magnitude and with the extent to which the degree of sensitivity to tuition charges varies depending on family assets and income, race, sex, and family background, academic aptitude, career interests, the kinds of schools in which a student is interested, and the amount and form of student aid that is available. There has been relatively little systematic analysis of these questions, and much more work is needed.[2] (In discussing the question of the strength of student response to changes in tuition, as reflected in application and enrollment decisions, one surprising tendency I have noted is for noneconomists to be more willing than economists to believe that economic considerations are likely to be very powerful)

Second, a form of money illusion can operate here as elsewhere, and economists can provide some perspective by relating changes in student charges to changes in money incomes There is a natural tendency for many people to see only the "bad" side of inflation, and thus to be aware of how much the prices of all kinds of things have risen (including the price of higher education) without recognizing that their own money income has also been going up. In the case of my own university, for example, many people are surprised to learn that between 1961–62 and 1974–75 student charges rose less rapidly than the nationwide level of median family income (102 percent and 124 percent being the respective rates of increase).

By encouraging us to see rising tuition in the context of nationwide trends in prices and incomes, I am not of course trying to suggest that there is no basis for concern about the potential effects of rapid increases in tuition. My own view is that there is serious ground for concern, especially if insufficient amounts of student aid are provided, and my third observation is that all of us need to work harder to help members

[2]One effort to relate application and enrollment decisions of individuals to a variety of economic, academic, and socioeconomic variables was made by Richard Spies in "The Future of Private Colleges: The Effect of Rising Costs on College Choice." Princeton University Industrial Relations Section Monograph, 1973 I am told that efforts are now under way to update and extend this study

of the executive and legislative branches of government see the social utility of ensuring genuine equality of opportunity and diversity of student bodies. I think it is important that financial aid programs be structured so as to provide students from low and middle income families with a reasonable opportunity to attend relatively expensive as well as less expensive institutions for which they are qualified (assuming, as I would, that the individuals should also be willing to share in defraying the higher costs).

The case is partly one of simple fairness. There are also, however, other important societal interests at stake. If our human resources are to be developed as fully as possible, financial considerations should not be allowed to prevent us from achieving the best possible match between the talents and aspirations of individuals, whatever their economic backgrounds, and the variety of educational opportunities offered by different colleges and universities. Also, educational objectives of no small consequence are served by having a student body composed of individuals from diverse socioeconomic backgrounds. Students learn in important ways from each other, and from the friction of differing perspectives and conflicting ideas. Moreover, in a society committed to the democratic ideal, there is a broad national purpose served by avoiding an economic segregation within higher education.

My fourth observation having to do with tuition, student aid, and related matters is that I hope we shall be careful not to base judgments concerning the appropriate levels of charges and of financial assistance on an overly narrow and predominantly economic definition of the benefits of higher education. A number of our colleagues have contributed importantly to the valuable body of literature on the measurement of returns to personal and social investments in higher education, and it is not, as a general rule, the authors of such studies who have failed to stress the importance of noneconomic benefits. Still, there is an ever-present temptation to exult what can be measured over what cannot, and it is a temptation that needs to be resisted.

To these rather general observations, which will be read by most people as having greater applicability to undergraduate than to graduate education, let me now add a brief postscript directed specifically to graduate education in the arts and sciences. While I certainly would not dismiss the relevance of projected needs for teachers in determining the appropriate overall level of investment in graduate education, particularly in some fields, I do worry about the tendency to think only in these terms, the tendency to assume that we know more than we can know about future supply and demand conditions, and the tendency to forget that both demand curves and supply curves have slopes.

There is an intimate interrelationship between graduate education and research in almost all fields, and it is essential that there be enough support for graduate education to provide continuity of effort and to insure that we not fail to educate those who will have to provide the academic leadership in the next generation. As I have argued elsewhere,[3] I believe that there is an established policy mechanism available, in the form of a program of portable graduate fellowships available to the most outstanding individuals on a competitive basis, which would distribute resources sensibly and which would enable us to meet our qualitative goals for graduate education without asking for unrealistically high subventions from the government.

III. A Sense of Community

In addition to worrying about the effects of financial pressures on the quality of higher education and on the size and composition of student bodies, I worry about effects on what I can call only our sense of community—on our shared commitment to common goals, on mutual respect, and on our ability to work effectively together on individual campuses and even across campuses. To be sure, these are intangible attributes—but I think we make a serious mistake if we downplay the importance of the right kind of milieu to our ability to serve well

[3]Remarks at a meeting of the Princeton Club of Washington, March 4, 1975

our educational and scholarly purposes. The "production function" of a university is peculiarly dependent on the attitudes of the all-too-fallible and all-too-human participants in the enterprise, and financial distress inevitably breeds some tension, suspicion, and often more than a little determination to protect one's own vested interest.[4]

I believe there are at least a limited number of organizational steps which can be taken to lessen the debilitating effects of these tendencies, but I am not going to trade on your patience today by trying to enumerate them. My purpose, instead, is simply to raise a large warning flag, to ask you to be aware of the heavy costs to our institutions of an excess of contentiousness.

Now it may be that in offering this comment I have fallen prey to another near-universal temptation: to ask the lion and the lamb to lie down together, especially if one is the keeper of the zoo. But then, I alerted you at the beginning to the obvious fact that I would be speaking from the perspective of the particular responsibilities that I happen to hold at present. Besides, when am I likely to have another opportunity to ask the members of this distinguished association to be kind to university presidents, as well as to students, staff members, alumni, and even their faculty colleagues unwise enough to have embraced different academic disciplines?

[4]Certainly the standard theory of the firm taught in elementary courses provides little sense of the importance of such considerations—which I believe to be particularly significant in organizations such as universities—not, I hasten to add, because I believe them to be inherently "inefficient," but rather because of the nonhierarchial nature of educational institutions with the accompanying stress on individual initiative and individual acceptance of responsibility to get on with one's own work and to help others. I do not know anything like enough about organizational theory to be familiar with efforts to study such phenomena systematically, but I am convinced of their importance.

ECONOMIC EDUCATION

What Economics Is Most Important to Teach: The Hansen Committee Report

By RENDIGS FELS*

My assignment was to write what amounts to a review article of a forthcoming publication by the Joint Council on Economic Education entitled "Framework of Basic Economic Concepts and Generalizations" (hereafter "Framework"). The assignment is unusual for two reasons. The publication is much shorter (about fifty typescript pages) than the book-length ordinarily deemed necessary to justify a review article, and the "Framework" is still subject to revision (possibly in response to this paper).

The "Framework" is Part I of the forthcoming *Master Curriculum Guide in Economics for the Nation's Schools* (hereafter *Guide*). Part II of the *Guide* will provide teachers of kindergarten and grades 1–12 (hereafter K–12) with detailed suggestions for how to teach the concepts and generalizations of Part I. Since Part II will build on Part I, the "Framework" is more important than its length might suggest, and economists have an obligation to see that it is as accurate and helpful as possible. The "Framework" has been prepared by a committee chaired by W Lee Hansen. The other members are G. L. Bach, James D. Calderwood, and Phillip Saunders. The committee has had the help of numerous other people.

The "Framework" is a successor to the *National Task Force Report on Economic Education in the Schools*, which was published in 1961. Like the *Task Force Report*, the "Framework" specifies economic concepts and generalizations deemed important for economic literacy. The "Framework" elaborates on a "reasoned approach" for analyzing economic policy issues which the *Task Force Report* had sketched but not emphasized.

The *Task Force Report* of 1961 was justly criticized for overambitiousness. Since it amounted to a summary of what every college student should know about economics after completing a one-year college course, the *National Task Force* proposals for what every high school graduate should know about economics were unrealistic. The "Framework" is a decided improvement over the *Task Force Report* in this respect. Its summary of concepts and generalizations is more modest. In another improvement, it makes some attempt to indicate priorities by attaching asterisks to the topics the Committee deems most important. Like everybody else who reviews the list of concepts and generalizations, I have a number of minor quarrels and complaints with it, which are detailed in the appendix to this paper. In addition, the list omits "other things remaining the same," which seems to me an indispensable analytic tool useful for all thinking, not just economics. Radical economists will, no doubt, have serious objections not only to the list but to the entire enterprise. But by and large, the list is one which orthodox economists can support. My principal complaint is that the concepts and generalizations are not stated with the high degree of precision and clarity that the purpose of the "Framework" calls for. Concepts are frequently introduced without being defined. When they are defined or explained, they are not always stated with complete accuracy and clarity. Since the "Framework" is to be used

*Vanderbilt University. The first draft of this paper included an appendix and some footnotes which have been deleted from the published version. They contained suggestions for minor revisions of the "Framework of Basic Economic Concepts and Generalizations."

by curriculum groups to prepare materials for use by teachers of K–12, complete accuracy and clarity, not just intelligibility to those who already know economics, is important. But the authors of the "Framework" have not done badly in this respect, especially in comparison with some authors of college texts who seem to think the way to make a difficult idea easy is to fuzz it up.

In his foreword, S. Stowell Symmes says the "Framework" "has taken a giant step beyond the 1961 [*Task Force Report*] because the authors have accepted the responsibility of putting down on paper some vignette of what it is like to *use* economics; that is, how economics becomes functional to thinking and deciding." I trust that this sentence will be revised before publication to say what the writer must have had in mind Setting down on paper some vignette of how to use economics hardly constitutes "a giant step." But what the "Framework" proposes could become a giant step in economics education if successfully implemented in the curriculum of grades K–12 It calls for putting major emphasis on training students to apply economic concepts not only to newspaper reports but also to reaching decisions on economic policy issues [1] It provides a "reasoned approach" (*RA*) to decision making consisting of six steps. defining the problem, selecting goals or objectives and indicating priorities, identifying the main options for attaining the goals, identifying economic concepts and principles useful for the problem, analyzing the likely consequences of each option, and deciding on the basis of the preceding steps which option is best.

At this point I must confess to bias The "reasoned approach" of the "Framework" is identical in substance, although a little different in organization and wording, with the standard operating procedure (*SOP*) for analyzing policy issues that I have been pushing for nearly a decade (see Fels and Robert G. Uhler). Since the organizer of this session is also a member of the "Framework" committee, getting me to evaluate the "Framework" almost looks like a put-up job. Inevitably I am delighted that the *SOP* is being incorporated into a report that will get far wider distribution and have much greater influence than I have been able to obtain for it. But perhaps because I am so thoroughly familiar with it, I am acutely aware of the magnitude of the task that the authors of Part I (the "Framework") are setting for the authors of Part II of the *Master Curriculum Guide* project. If the present committee is less ambitious and more realistic than the 1961 Task Force with respect to concepts, it is more ambitious and less realistic with respect to policy analysis.

Internal evidence within the "Framework" report itself supports the twin fears that training students in grades K–12 to use the reasoned approach is a formidable undertaking and that the committee has underestimated its magnitude. The "Framework" contains one section of text and an appendix apparently intended to demonstrate how the reasoned approach should be used. The appendix is entirely useless for the purpose It consists solely of a reprint of an article from the *Conference Board Record* by Michael H. Moskow called "Environmental Regulation and Public Values." The article itself is excellent But the connection between it and the six steps of the "reasoned approach" has not been made. I like to think I could work out the connection if I were to spend a few hours trying, but the people for whom the "Framework" is intended—curriculum specialists and teachers of grades K–12—can hardly be expected to do so The fact that the committee has not provided the connection leads to a darker suspicion which I shall return to later.

Section X of the "Framework," which is entitled "Applying the Elements to Particular Issues: Some Illustrations," is somewhat more helpful than the appendix just referred to, but it stops well short of demonstrating how students in grades K–12 or their teachers are expected to use the reasoned approach. It provides a series of five newspaper headlines on coffee prices together with references to the economic concepts included elsewhere in the report that are

[1]At least I think the committee intended to put major emphasis on such training. There is some inconsistency in "Framework" which is discussed in the appendix to this paper.

useful for understanding what the headlines are all about. But it does not provde the connection between the concepts named and the headlines. Readers are apparently expected to work out the analysis for themselves. Similarly, there are only brief comments on how students are expected to go about making up their minds on the policy issue raised by the last of the headlines. If we economists are to provide help to curriculum specialists and teachers of K–12, we have to do better than this.

Such shortcomings in the version of the "Framework" report now available can and probably will be corrected before it is published. But they are symptomatic of a more fundamental problem They would not have shown up in the semifinal version of a report by a group of distinguished economists if the task of applying economic concepts to headlines on coffee prices and of using the reasoned approach to analyze policy issues were not considerably harder than the committee seems to think I do not doubt that the committee could do the job. The fact that it did not do so suggests that even the committee does not find it all that easy Curriculum specialists and teachers of K–12 will find it so much harder. And the students . . . ?

This calls to mind a searching question put to me in the 1960's by my old friend Alice Bourneuf. As a visiting speaker at Boston College, I had made my pitch for training college students in the elementary course to analyze policy issues for themselves. "But," she asked, "can they do it?" I have thought of answers, but the question has gnawed at me ever since.[2] I now have evidence that college students can be trained to use the reasoned approach, but the evidence also indicates that mastering it takes considerable time. This implies that training teachers of grades K–12 to use the reasoned approach will be a major undertaking.[3]

Training the teachers would be a major undertaking even if enough economists were available who knew how to provide the training. But there are not. The "Framework" committee's reasoned approach is only beginning to creep into textual materials for college instruction. Professional economists do not normally use it in their teaching or in their writing. This is true even though the reasoned approach is nothing but an adaptation of a procedure pioneered by Jan Tinbergen which is increasingly being used in econometric analysis of policy issues. To borrow a line of Robert Solow's, the *RA* may be the wave of the future but it is not the wave of the present. Since very few professional economists use the reasoned approach themselves, there are correspondingly few with experience and skill in training others to use it.

Bourneuf's question has another dimension. What can we realistically expect high school graduates to be able to do in the way of analyzing economic policy issues for themselves? Even if their teachers have mastered the reasoned approach and even if high quality curricular materials become available in abundance (a problem I shall comment on later), do we really expect high school graduates to read articles like the one by Moskow or headlines like the ones on coffee prices and then analyze the issues for themselves in a way that professional economists do not commonly do? I fear not. I propose a more modest goal. Give the high school graduate two articles, both in the format of the reasoned approach, which reach different conclusions on a controversial issue. The graduate should be able to choose between the two and articulate sound reasons for the choice.

[2]The question, of course, is not whether they can do it but how well they can do it, and whether giving them practice in college courses has a social product greater than the alternative use of the time One answer is that students are bound to have opinions on economic issues, and their opinions will influence political decisions We cannot afford to say that economic policy is too difficult for ordinary people and should be left to professional economists. Any training we offer is likely to lead to students doing better. Whether the degree of improvement warrants the opportunity cost of omitting, say, isoquants from the elementary college course is a question of both reality judgments (for which evidence would be desirable) and value judgments. In my opinion it does.

[3]Four-week workshops for teachers similar to the numerous ones now sponsored by the various state Councils on Economic Education could be used. Four weeks devoted entirely to giving teachers practice in using the *RA* on a series of economic policy issues may suffice provided the teachers already knew some economics.

This leads to another problem, the lack of articles in the format of the reasoned approach. To train students in the *RA* will require teaching materials at the intellectual level appropriate to grades K–12. At present they simply do not exist. There is no pool of articles that can be drawn on or readily adapted for the purpose. They will have to be worked up from scratch. The difficulty of doing so is illustrated by the unsatisfactory attempt of the "Framework" committee to provide a single example based on headlines about coffee prices. Likewise the absence of the reasoned approach in newspaper articles, political speeches, and the writings of economists means that today's high school graduates would not be able to use the skill I have proposed as a realistic goal because they would not encounter anything in later life to use it on. Until the reasoned approach permeates national discussions of economic policy, even the modest educational goal I have proposed has only limited value for attaining what the "Framework" committee regards as the goal of economics education, responsible citizenship

In spite of all these difficulties, I want to urge the "Framework" committee and the curriculum specialists who will build on its report to persevere with the reasoned approach. For the *RA* is just what its name says it is, a reasoned approach to economic policy. It is more than that. It is a useful format for *any* decision making on any kind of problem involving multiple objectives, whether economic or noneconomic, whether social or personal. As such, its educational value is potentially high Because it is a framework for rational analysis, its use in political debates and newspaper discussions would improve the quality of national consideration of issues of political economy, possibly leading to better policy decisions by the government. For the same reasons, we can expect, or at least hope, that its use will continually increase until it becomes as common a part of the intellectual equipment of educated people as the multiplication table. If so, work on introducing it into grades K–12 should begin at once. By the time doing so could, on the most optimistic expectations, have a significant impact on the economic literacy of high school graduates, there may well be an abundance of newspaper articles on which they can use their skill.

REFERENCES

Rendigs Fels and Robert G. Uhler, eds., *Casebook of Economic Problems and Policies: Practice in Thinking*, St. Paul 1974, 1975, and 1976.

Michael H. Moskow, "Environmental Regulation and Public Values," *The Conference Board Record*, April 1976, 47–50.

Jan Tinbergen, *Economic Policy and Design*, Amsterdam and Chicago 1967.

Economic Education in the Schools, Report of the *National Task Force on Economic Education*, Committee for Economic Development, New York 1961.

Joint Council on Economic Education, "Framework of Basic Economic Concepts and Generalizations," review edition (mimeo.), New York 1976.

Teaching Principles of Economics: The Joint Council Experimental Economics Course Project

By ALLEN C. KELLY*

For the past several years the Joint Council on Economic Education has engaged in a project to identify and assess alternative approaches to the teaching of college introductory economics. The goals of the project, as summarized by Arthur Welsh, have been " . . . to develop alternative approaches that overburdened professors in two- and four-year colleges might find more useful than their current offerings and to encourage others to improve and expand upon the Joint Council's efforts" (Rendigs Fels, p. 1).

Several professors and schools have participated in this effort: Kenneth and Elsie Boulding of the University of Colorado, Rendigs Fels of Vanderbilt University, Richard H. Leftwich and Ansel M Sharp of Oklahoma State University, Phillip Saunders of Indiana University, and Barbara and Howard Tuckman of Florida State University. Syllabi and supporting materials have now been published as special issues of *The Journal of Economic Education* for the last four of these courses (Fels, Leftwich and Sharp, Saunders, Tuckman).[1] The course developed by the Bouldings was reported upon in preliminary form at the December 1973 American Economic Association meetings, and thus I will concentrate my attention in this review on the remaining four.

I. Overview and Comments on the Individual Course Packages

There is a common set of premises underlying the four courses which have been developed. Fels summarizes these premises well. "Standard textbooks are typically overloaded. All too often instructors feel obliged to assign the whole book, leading to overloaded courses. As a result, the student gains vague familiarity with a wide range of economic theory and a mastery of none of it. In addition, the typical course provides no training in the skills of applying economic principles" (Fels, p. 5).

The course developed by Fels carries this theme to its most extensive level. This course is structured around teaching the *application* of economic theory to a wide range of realistic problems. Students review "cases" which take the form of relatively short expositions of economic situations—often based on quotations from newspapers, or on edited newspaper articles. Students then work through a carefully constructed set of questions designed to train them in the process of orderly thinking about economic problems. To make the teaching approach more easily adopted by others, a casebook and instructor's manual have been developed, coauthored by Fels and Robert G. Uhler. Fels's syllabus shows in detail how cases may be incorporated into a "conventional" course (i e., the traditional lecture format), identifies the key concepts which should be drawn out for each lesson, and even provides sets of notes to aid professors in leading class meetings.

In addition to the case application emphasis, Fels has developed his course around the Personalized System of Instruction (*PSI*) approach pioneered by Fred S. Keller. *PSI* employs virtually no lectures. The lecture period (if used for formal instruction) can be devoted to other activities: discussion, test taking, tutoring, projects, and so forth. Students may proceed at their own pace. Their performance is evaluated on the number of course units mastered. Assessment of mastery learning over the various in-

*Professor and Chairman, Dept of Economics, Duke University I am grateful for the comments of W. Lee Hansen, John J Siegfried, and Sue Whitesell on an earlier draft of this paper.

[1]To obtain a copy of the four syllabi, write Publications Department, Joint Council on Economic Education, 1212 Avenue of the Americas, New York, NY 10036

structional units is accomplished by written or oral examinations. Students have considerable flexibility in selecting the time to take the examinations; these may also be retaken any number of times, until mastery achievement has been demonstrated. Because of the extensive amount of testing required in such a course, student undergraduate proctors are used. The proctors are also available for student consultation and tutoring. They receive three credits toward their economics major for participating as proctors. Budget constraints will typically preclude using exclusively higher-paid graduate students in this role. There is one proctor for approximately ten students. Ten proctors is around the maximum a professor can supervise; thus, *PSI* classes do not usually exceed one hundred students.

We are fortunate to have available, even at this early state, an evaluation of Fels's course by John J. Siegfried and Stephen H. Strand, neither of whom was involved in the course development or in its implementation (Siegfried and Siegfried and Strand). Their careful studies reveal that

> ". . . (1) *PSI* students performed no better or no worse on multiple choice or essay examinations than students in the conventional lecture course; (2) there was no difference in performance in subsequent economics courses . . . , (3) students liked the course more, thought they learned more, and felt they were examined and graded fairer in the *PSI* course; (4) there was no difference in the amount of time spent on the course activities between *PSI* and conventional course students; (5) the tendency to elect economics as a major was unrelated to the method of instruction . . ., and (6) the student-proctors learned more economic theory than they would have learned from an alternative upper class economics elective." [Siegfried, pp. 32–33]

The latter effect, while based on a small sample, is quantitatively quite large.

Fels's contributions lie in two quite separate areas: the application of *PSI* to economics, and the development of the case-application approach. In a sense he has provided two course packages.

Of the four course packages under review, Fels's course is not only the most innovative, but also the most complete and diffusible. It includes the course syllabus, illustrative examinations, notes to professors to guide their discussion sessions, cases, instructions to students, and so forth.

A few relatively minor qualifications might be expressed relating to Fels's course. First, the course is highly labor intensive—especially the *PSI* format, but to a lesser extent the case approach. To obtain proficiency in economic analysis, students must write several cases; these must be graded, and students must be provided detailed feedback. While I am convinced that the case-applications approach represents a superior way of teaching the most important elements of economic principles, I would still like more information on alternative formats of using this approach which employ less labor inputs. I suspect that a totally case-oriented course with active student involvement, and a course which is also economically feasible in a wide range of colleges and institutions, is yet to be discovered. I hope that considerable experimentation with alternative formats will be stimulated by the high quality materials made available by Fels and Uhler. Equally important, I hope that professors evaluate the effectiveness of these alternative formats, and then report their results.

A second qualm relates to the difficulty of the materials. While Fels has provided suggestions for lowering the difficulty level, I doubt if these suggestions will be sufficient. He has set his standards high. More experimentation will be required to develop a less demanding set of course objectives—possibly by trimming some of the content, but not necessarily the level of required analysis. Parenthetically, some would seriously question the feasibility of the goal of teaching freshmen and sophomores how to engage in simple and complex application of economic principles. My own position is that this goal may not be feasible for the majority of this group of students. However, the value of the benefits obtained for those who achieve this level of proficiency will far exceed any foregone learning of "tools" usually taught in the principles course, since these tools, typically taught without extensive application, are rapidly forgotten for lack of purpose.

Finally, it must be kept in mind that the *PSI* approach, while promising, is costly. Fels has

noted that course instructors should probably be awarded more than one course "credit" to provide them the incentive and time to implement *PSI*. Moreover, critical to this approach is the availability of student proctors. It is extremely encouraging that Siegfried's research has shown that the benefit of proctoring to the learning of economics notably exceeds the opportunity cost of the typical upper division economics course. More hard evidence on this point from other studies could be decisive in making college administrators and faculty receptive to providing course credit for proctoring, and thus effectively opening up the financial and technological viability of *PSI*.

The course by Leftwich and Sharp centers around the "issues approach" to teaching economics, and also stresses application of economic principles. They have developed a book which focuses on the various issues taken up in their syllabus. They have also provided references to several other books which offer pertinent cases and materials. Their issues are generally broader than Fels's cases, and as a result, there are fewer of these issues, and each one occupies around a week of the course. The syllabus itself provides for each issue "major discussion points," "economic concepts and principles" covered, and a "recapitulation." While these aids are useful, the course package would have been considerably more valuable and adaptable by others had the authors also provided guidelines and suggestions to the professor for actually organizing and leading the various discussion sessions, together with test items which evaluate the students' mastery of the various issues. The issues, however, are well chosen and should engender considerable student interest. Moreover, in unpublished research results, the authors have argued that the exclusive teaching of economics around these several issues does not result in any notable sacrifice of the basic tools typically taught in the standard economics course.

Saunders' course confronts the difficult problem of developing and coordinating an application-oriented offering involving the participation of as many as twenty different instructors. He has assembled with the participation of his colleagues a concensus on a required "core" of economic analysis which is common to all sections, and which is tested by a common final examination. Each professor is then free to develop specific application emphases, e.g., environmental economics, current economic issues, income distribution. Students are free to select the orientation that interests them most. A review of the alternatives open to students at Indiana University reveals a smorgasbord of courses seldom found in even the combined offerings of a dozen institutions of higher education.

Saunders' contributions lie mainly in his interesting and useful presentation of information on course planning, teaching techniques, and research formulation. He has presented valuable tips, in highly readable form, on such topics as the formulation and evaluation of examinations, the development and assessment of course evaluations, the preparation of course objectives, the construction of research designs for appraising teaching, the techniques of coordinating courses with many professors, the alternative ways to use and train graduate student instructors, and so forth. Some useful base-line data are also presented on course evaluation surveys and examinations.[2] In my judgment, his syllabus is most helpful on the general methods of course development, and especially on the techniques of coordinating and influencing many instructors to adopt a common core of economic analysis, yet allowing each professor to "do his or her own thing." This format makes faculty more amenable to undertaking the perceived "chore" of teaching principles of economics, and at the same time, increases their effectiveness.[3]

[2]For example, on a point that has received considerable attention (*J. Econ Ed.*, Fall 1973), Saunders' data collected over eight semesters from almost nine thousand students indicate that student performance on the common final examinations is positively and significantly associated with the evaluation rating of the instructor from whom they took the course

[3]Subsequent to the publication of the syllabus being reviewed here, Saunders has developed two student workbooks—one for microeconomics and one for macroeconomics—that define the analytical core for each semester of Indiana's introductory course in terms of specific examination questions and a set of 15–20 homework problems. A special seminar has also been developed to train graduate student instructors to become more effective teachers of introductory economics.

Barbara and Howard Tuckman's course is possibly the most traditional of the ones offered, and as a result, will be quite easily adopted by many institutions. Its innovations are less in the course content area, and more in the areas of providing techniques for motivating students to learn economics, and providing students flexibility in their pace of studying economics and of taking examinations. Students are placed in a simulated environment of being actual consultants to the government and are required to complete "memoranda" (disguised workbook and/or case application-like exercises of a policy orientation) for the President and his key advisors. The authors also offer a few specific and useful ideas for arousing student interest during the classroom presentation of some key economic tools: the consumption function, the multiplier, and so forth. Their attempts at self-pacing of student learning met with mixed results. When provided freedom of time allocation students procrastinated, creating administrative and testing burdens at the conclusion of the course. The authors have thus elected to offer rewards and penalties (in the form of course points) which have the net effect of bringing many of the students back toward the traditional study mode in terms of time allocation. Their findings accord with my own hunch that the self-paced mode is suitable primarily for the more highly motivated and disciplined students—and these, lamentably, are not the vast majority of students in American higher education.

II. Overall Appraisal of the Joint Council Program

With the exception of Fels's package, each of the remaining syllabi is somewhat incomplete in providing all the elements of a "turn-key" offering for the overburdened economics instructor. However, I do not consider this a telling deficiency of the developmental effort. Something should be left for the instructor to do, overburdened or not More important, I find that the various packages, *when taken together*, provide a surprisingly wide variety of materials and ideas that a large number of instructors should find useful in their course planning and implementation. For example, the instructor will find very helpful Saunders' and Tuckmans' techniques of course planning and course coordination with many teachers; the instructor may utilize some of Fels and Uhler's cases, and some of the issues provided by Leftwich and Sharp; and the instructor will also hopefully engage in some course evaluation of the type illustrated by Siegfried, Strand, and Saunders.

Other conspicuous achievements of the Joint Council program are both the respectability it has provided the idea of teaching more limited course content with higher emphasis on application, and its many suggestions on ways for accomplishing these goals. Two books containing useful case and issue materials have resulted from the project; others are now available from commercial publishers, in part in response to the project itself While the optimal format for using the case approach may not yet have been discovered, this approach is here to stay, and will gradually improve over time.

Finally, these syllabi will serve to stretch the thinking of the teacher trapped in the routine of providing students with a comprehensive coverage of encyclopedic texts. Such instructors, with the aid of these Joint Council materials, can rediscover some of the excitement of teaching For example, while the Tuckmans' Presidential Policy Memoranda do not represent a notable breakthrough in concept coverage, this technique does indeed illustrate an innovative way of "packaging" economics to engender some increased enthusiasm and interest—and that contribution is not to be discounted. These syllabi will encourage other professors to break out of their traditional mold, and to discover new techniques of teaching economics in a more exciting and effective manner. Certainly the syllabi demonstrate conclusively that this field is wide open, and that this activity can be intellectually rewarding.

In pointing to the future of the experimental courses under review, and to other developmental efforts of this type, two broad issues should be raised. First, it would be good to establish a modest program to track over time the successes and failures of these courses. One might, for example, document over a period of

five years the experience of a group of adopters, asking each to respond periodically to a carefully constructed series of questions. Several issues come to mind. How many of the original set of adopters continue to use the materials after five years? Why do they continue (or stop) using the materials? What role do the adopters play in changing the original course packages, and in what specific ways?

Second, a review of these syllabi raises anew the critical question of how best to evaluate and compare various courses and teaching approaches. While instruments such as the Test of Understanding in College Economics are useful, this test is not by itself sufficient to compare such divergent courses as those represented by the syllabi under review. The time is ripe in economic education research 1) to identify a fairly comprehensive list of outputs which are likely to be impacted by changes in course content and teaching technique, 2) to develop reliable and standardized instruments and procedures for measuring these outputs, and 3) to attempt to obtain some weights on the value of the outputs as provided by the various clientele of our teaching programs Such a "Manual of Instructional Outputs and Their Measurement" would include items ranging from student enrollments, the number of majors, student "enjoyment" of the course, faculty willingness to teach the course, to such items as student problem solving skills in general, knowledge of economic tools, ability to apply economic tools, social and political values, and the like[4] The ability to utilize the wide range of research results that have been forthcoming in the *Journal of Economic Education* and elsewhere is increasingly being constrained by the heterogeneity of the set of instructional outputs chosen by various researchers to evaluate their teaching, and the wide range of techniques employed to measure even quite similar outputs. Comparisons are therefore notably hampered. We should begin thinking about *systematizing* our efforts in identifying, measuring and weighting the various outputs of education in general, and of economic education in particular. A careful evaluation of the Joint Council experimental courses could represent an excellent application of, and stimulus to, such an outcome.

[4]Attention to developing and systematizing such a set of instruments will be provided in a recently undertaken five-year project founded by the National Institute of Education, and undertaken by Richard Attiyeh, Keith Lumsden, and myself We will be developing several alternative course packages for teaching economics, employing Teaching Information Processing System, programmed learning, computer games, cases, and the conventional lecture technique Several dozen schools will be participating in the testing and evaluation of these various packages. The ability to identify, measure, and compare instructional outputs across course packages is clearly critical to the research

REFERENCES

Kenneth and Elise Boulding, "Introducing Freshmen to the Social System," *Amer Econ. Rev. Proc.*, May 1974, *64*, 414–19.

Rendigs Fels, "The Vanderbilt-*JCEE* Experimental Course in Elementary Economics," *J. Econ. Ed.*, Winter 1974, Special Issue No. 2

——— **and Robert G. Uhler,** eds , *Casebook of Economic Problems and Policies: Practice in Thinking*, St. Paul 1976.

Richard H. Leftwich and Ansel M. Sharp, *Economics of Social Issues*, Dallas, 1974.

——— **and** ———, "Syllabus for an 'Issues Approach' to Teaching Economic Principles." *J. Econ. Ed.*, Winter 1974, Special Issue No. 1.

Phillip Saunders, "Experimental Course Development in Introductory Economics at Indiana University," *J. Econ Ed.*, Fall 1975, Special Issue No. 4.

John J. Siegfried, "Is Teaching the Best Way to Learn? An Evaluation of the Benefits and Costs to Undergraduate Student Proctors in Elementary Economics," *Southern Econ. J.*, January 1977.

——— **and Stephen H. Strand,** "An Evaluation of the Vanderbilt-*JCEE* Experimental *PSI* Course in Elementary Economics," *J. Econ. Ed.*, Fall 1976.

Barbara and Howard Tuckman, "Toward a More Effective Principles Class: The Florida State University Experience," *J. Econ. Ed.*, Spring 1975, Special Issue No. 3.

CAPITAL FORMATION: WHERE, WHY, AND HOW MUCH?

Capital Shortage: Myth and Reality

By ROBERT EISNER*

A couple of years ago a New York Stock Exchange study (1974) pointed to a "capital shortage" of some $650 billion by 1985. Treasury Secretary William E. Simon, comparing his estimates of capital requirements in current dollars over the next decade with capital expenditures in current dollars over the last decade, came out with a gap of over 2-1/2 trillion dollars without noting the noncomparability of prices (p. 3871)

We have indeed a host of estimates from a number of econometric models, government bodies and private institutions, from Barry Bosworth, James Duesenberry and Andrew Carron and many others A major Bureau of Economic Analysis study under the direction of Vaccara projected a total of $986.6 billion, in 1972 prices, for business fixed investment from 1975 to 1980, or 12.0 percent of cumulative gross national product, "in order to insure a 1980 capital stock sufficient to meet the needs of a full employment economy, and the requirements for pollution abatement and for decreasing dependence on foreign sources of petroleum" (p. 7).

Scarcities are sometimes seen in terms of sources of financing. Benjamin Friedman wrote in 1975, "To an unusually great extent, financial considerations may act during this period [1977–81] as effective constraints on the amount of fixed investment which the economy in aggregate is able to do" (1975, p. 52). In May 1976, however, Allen Sinai declared, "There are no financial shortages of any consequence" (p. 2)

But with the plethora of articles, studies, claims and warnings, what meaning can we attach to the notion of a capital "shortage"? In what sense can there be a shortage in a free economy where markets are cleared by the impetus of price movements? In an uncontrolled, competitive system, the rate of investment is not imposed as a prior constraint. Business investment, in particular, is the resultant of the utility-maximizing saving propensities of households and the profit or wealth-maximizing production decisions of business. These are subject to the constraints of the general economic atmosphere determined by the monetary and fiscal authorities of government, particular tax and monetary influences, and general currents of the world.

Any argument that there is a capital shortage must either imply a literal failure of market clearing or some standard external to the economic system A failure of markets to clear in an equilibrium sense implies fixed or sticky prices If government were to control prices and set those for capital goods too low, the quantity of capital goods demanded could exceed the quantity of capital goods supplied. Perhaps more to the point, government regulatory agencies might hold prices of certain products, such as electric power, so low that, while the quantity of electric power demanded might be very high, firms anticipating continued low prices would not find it profitable to invest in the capacity to meet future needs.

Similarly, there may be price fixing in financial markets. If the monetary authority and/or inflation force up interest rates while regulatory

*William R Kenan Professor of Economics, Northwestern University, and Senior Research Associate, National Bureau of Economic Research. I am indebted to Martin Feldstein, Benjamin Friedman, Marc Nerlove and Beatrice Vaccara for helpful comments

agencies offer restrictions on what interest may be paid, various kinds of shortages may develop. In some instances regulatory requirements of earnings coverage on debt issues may make impossible further corporate borrowing. At the same time, investor expectation of future returns may be such as to make the cost seem prohibitive for raising funds through sale of additional equity. Restrictions on interest rates paid by various banking and nonbank lending institutions may also have the effect of drying up the supply of funds for certain kinds of investment, particularly for residential construction which traditionally looks to such regulated institutions for financing.

Curiously, most discussions of alleged capital shortages do not focus sharply on these particular interferences with the free functioning of product or capital markets. Neither do they point rigorously to positive externalities of private saving and investment or negative externalities of current consumption, private or public, which might warrant government intervention in these markets in support of capital formation. Rather they relate to imagined disparities between the amount of capital or the rate of investment which some individual or group asserts we *should* have and what appears to be forthcoming. On the real side, projections are made of future rates and composition of production, levels of employment and the amount of capital "required" at some specified future date to match the given employment and output. Some judgment is then made as to whether the rate of saving over the intervening period will be such as to accumulate a sufficient amount of capital or what governmental policies might be appropriate to bring about such saving and investment.

As probably the most meticulous, thorough and detailed estimate of business fixed investment "requirements," the Vaccara-*BEA* study permits us to view clearly the basic inherent deficiencies of use of such projections to document a capital "shortage." First, the Vaccara-*BEA* work uses an extraneous Bureau of Labor Statistics estimate of 1980 "full employment" *GNP* and a sectoral composition of that *GNP* which predetermines the proportions of gross national product devoted to more and less capital-intensive final demand vectors. Second, capital-output ratios are determined from historical figures, sometimes with projections of trends in these ratios. No adjustment is made for the effects of possibly changing interest rates, prices, availabilities or costs of obtaining capital. Third, "summary" assumptions are made about discards or retirements and consequent need for replacement. No adjustment is made for the possibility that, faced with "shortages," firms might discard existing plant and equipment less rapidly. Fourth, requirements for pollution abatement capital are taken from *BEA* and McGraw-Hill projections and "a large dose of judgmental adjustment." Fifth and finally, needs for energy-related investment are taken from "Project Independence" programs.

Out of all that came the estimate of $986.6 billion as additional capital needed from 1975 through 1980 to meet the projected expansion needs for the specified final product mix in 1980 with also specified capital-output ratios and discards or retirements. To relate this to a projected flow of saving, real or financial, and infer a capital shortage would be to put economic processes in a strait jacket. If the indicated saving were not forthcoming at existing rates of return, would not the return to saving and the cost of obtaining it rise? Would not discards and retirements slow in the face of more costly capital? Would not industry shift to less capital-intensive or less durable means of production, thus reducing capital-output ratios? Would not demand and the final product mix, under the pressure of changes in relative prices, shift toward less capital-intensive industries? And might not the market output to be produced by a full employment economy be reduced in response to the shifts in allocation to the nonmarket output of pollution abatement or to more costly domestic energy production?

While much business attention is directed to presumed shortages in the financing of business investment, it is hard to believe or to find in the

data evidence that our financial system is unable to complete the nexus between savers and the accumulaters of real capital. As we have suggested, imperfections in our financial markets, frequently created by government restrictions, may well distort the allocation of saving. Certain restrictions, such as those on interest payments on deposits may to some extent discourage saving, but even here conclusions depend upon the doubtful elasticity of saving with respect to its rate of return, including both income and substitution effects.

Individual firms at times believe themselves pinched by financial shortages in the face of what appear to them to be attractive investment opportunities. But in any economy where resources are not free, there are opportunity costs to investment Costs to an individual firm, financial and nonfinancial, reflect market valuation of alternative uses of desired resources If an individual firm finds that it cannot obtain funds at a sufficiently low cost to warrant their use in investment, this in principle implies that there are other uses of those funds which are deemed more valuable

Where, in the aggregate, firms feel that they cannot profitably finance as much investment as they wish, households, nonprofit institutions and governments and government enterprises apparently have exercised superior claims to the additional resources which business might elect to have for more investment. This, ultimately, is not a financial constraint but a real constraint imposed by the limitation of resources on the one hand and society's preferences, expressed both individually and socially, on the other.

The decisive constraint on capital formation may well lie in the supply of saving, although not in the manner sometimes affirmed. "Gross saving" in our national income and product accounts comprises personal saving, undistributed corporate profits, business capital consumption allowances, the government surplus and net capital grants received by the United States. This is identically equal to gross investment, which includes gross private domestic investment and net foreign investment. The identity is a powerful and sharp but potentially misleading tool, where one is tempted to apply carelessly *ceteris paribus* assumptions. One might, for example, assert that, given gross saving which equals gross investment, reducing net foreign investment would raise gross private domestic investment. But can one properly assume that reducing net foreign investment, with likely consequential reductions in the domestic employment and income associated with the production of goods and services sold abroad, would leave gross saving unaffected?

A most common complaint is that the federal government budget deficit, calculated at $74.6 billion in the 1975 National Income and Product Accounts, is "crowding out" private investment. We should, at least in this context and indeed more generally, dismiss the monetarists' argument that funds used to buy federal debt are not available to buy business debt. For this quite confuses stocks and flows of funds and fails to recognize that the money used to buy federal securities is in turn, roughly to the extent of the deficit, respent and hence again available for further lending. All this may create some pressure on interest rates if the monetary authority is not accommodating but even apart from that "if," there is no reason to anticipate major interest effects on investment.[1]

In terms of the saving-investment identity, what of the argument that of the $262.8 billion of gross private saving in 1975, $64.8 billion was dissipated in the government deficit (negative surplus, with a $9.8 billion state and local surplus partially offsetting the federal deficit)? It can be stated that only $195.4 billion was left for gross investment. Would not gross investment have been more if the government deficit offset to gross private saving were less?

Again such reasoning involves invalid *ceteris paribus* assumptions. Suppose the federal budget deficit were reduced by eliminating revenue-sharing grants to state and local governments. Would that not reduce the state and local governments surplus? Or suppose social

[1]A paper by Patric H Hendershott (1976) points out that a deficit-creating tax cut accompanied by increased short-term Treasury financing may well *lower* the long-term interest rates most relevant to investment

security benefits were reduced or personal income taxes increased. Would this not reduce personal saving? Or suppose corporate profits tax rates were raised. Would this not reduce undistributed corporate profits?

Even merely within the accounting framework, one quickly sees that reducing the federal budget deficit in an effort to make more private saving available for business investment may merely reduce other components of gross saving, leaving no more for investment. The full economic consequences may indeed be perverse. It should be clear to most that in a year which witnessed the depth of the sharpest and most severe recession since the Great Depression of the 1930's, action to reduce the government deficit, either by increasing taxes or reducing government spending, could only have been expected to further reduce aggregate demand, income, output and private saving. That recession saw the total of fixed investment drop 25 percent from the first quarter of 1973 to the second quarter of 1975 Any further government contributions to lowering actual demand by attempted budget balancing could only have depressed the economy and saving and investment all the more

If some future capital "shortage" is foreseen, the surest and most substantial spur to current capital formation is a rapid return to relatively full production and employment. There need be no fear of lower taxes stimulating consumption or increased government spending depriving capital goods industries of resources when unemployment and excess capacity are rampant.

Once we contemplate full employment, the rules are quite changed. With resources fixed in the short run, in any economic world we know there are scarcities everywhere Households would like to consume more. Those concerned with the provision of public goods—or whatever else comes from government—would like more of them. And those responsible for the acquisition or production of capital to meet future needs would like to have more of that. Who is to say that there is to be a greater allocation of resources to one of these categories—the accumulation of capital—and less to the others? "Shortage" becomes merely the somewhat pejorative expression of the universal characteristic of scarce resources.

One argument for the existence of a "capital shortage" is that government policy, particularly tax discrimination, has biased capital accumulation downward. With regard to business investment, where most of the heat has been generated, such an argument is not easily substantiated. Rather, the combination of capital gains exclusions, tax depreciation in excess of economic depreciation, tax deduction of interest costs, and equipment tax credits, particularly in a climate of expected inflation of capital goods prices, offer a considerable distortion in the direction of more business investment than would be undertaken in a free market. This is probably accentuated by complementary restrictions of investment in housing, government, nonprofit enterprises and human capital.

It is indeed in these latter categories that we may find greatest evidence of true capital shortage Anticompetitive forces in the area of building trades and residential construction, along with restrictive covenants and imperfect mortgage markets, may well be accountable for depressed investment and excess capacity in the home building industry Government military expenditures receive vast support, but a systematic effort to decide on public investment in terms of cost-benefit analyses, which would correspond to entrepreneurial profit calculus, might give different results from those stemming from the current electoral-legislative-logrolling complex. Neither nonprofit enterprises nor state and local government, we should be reminded, enjoy any benefits from equipment tax credits or accelerated depreciation.

But most important is the great bulk of capital accumulation which takes place in intangible or human form. Here there are basic a priori reasons to expect underinvestment. Where a company constructs or buys plant and equipment it can retain it and its benefits for itself. Where it invests in research, development, know-how and training, since knowledge and skills are generally freely disseminated in a free

society, differences may be substantial between marginal return to the investor and marginal social return. Most particularly, since we are not a slave society, it does not pay individual private enterprise to invest in human beings for more than the expectation of returns from their uncertain and usually short-run employment.

Yet the serious imperfection in human capital markets, along with understandable individual risk aversion, makes it very difficult for people to invest adequately in themselves. Information and transaction costs curtail drastically the supply of finance for human capital. What youth with aspirations for business leadership or service as an engineer, political leader or economist can go to the bank and say, "Invest in me! My expected life-time earnings are high. I would be happy to give you a promissory note or sell you equity rights in my human capital"?

As Benjamin Friedman has suggested (1976), the issue of capital shortage may perhaps better be raised as. Shortage for whom? The sometimes heated discussion may have more to do with distribution of income and particularly wealth than with their aggregates. Tax concessions to business allegedly to encourage investment essentially convey ownership of additional capital to current equity holders. General cuts in taxes to stimulate demand and indirectly encourage investment give increased capital ownership to all those who save more out of increased after-tax incomes. Expenditures for education and training increase the wealth primarily of those whose only capital is human.

Finally, it is argued that government transfer payments and taxes create a capital shortage in the sense of encouraging consumption and discouraging saving. In part this argument depends upon notions, appropriately questioned in Milton Friedman's permanent income and Franco Modigliani's life cycle consumption functions, that the marginal propensity to consume of the poor is greater than that of the rich, so that redistribution from the rich to the poor will raise consumption. Indeed, the dominant component of taxes on the working young to finance transfer payments to the elderly retired may suggest quite the opposite. The propensity of Americans to leave estates may be such that, despite the need of many elderly to consume all of their social security benefits, our social insurance system may add more to private saving than it subtracts.

Concerns that the social commitment to retirement benefits vitiates the need for and hence reduces the quantity of private saving may be countered on two counts. First, they ignore the effects of alternative private commitments, chiefly from one's children. Second, they raise some question as to the appropriate arguments of a social welfare function. If people prefer to avoid risk and uncertainty as to their retirement and to avoid having to save to meet that risk, why should government not permit them to obtain this superior position?

It is also asserted that a capital shortage is created by income taxation which reduces the after-tax return on saving. But here we must keep in mind both income and substitution effects. If saving is motivated by expected future consumption needs, a lower rate of return on accumulated wealth may induce us to save more in order to reach or come close to our originally preferred consumption path. The same argument of course applies to the effects of taxation on productive or remunerative work itself. As taxes rise we have to work more to attain any given level of after-tax benefits.

Finally, we are told that, for some reasons of state or religion, we must accumulate capital more rapidly in order to grow faster. It is argued that alleviating capital "shortage" would contribute to growth and hence to future output. But this would be at the expense of current availability of private and public goods and services. Is it necessarily desirable that we have more in the future than in the present? It is not axiomatic that we should sacrifice more when we are young in order to live better when we are older, or that our generation should sacrifice in the prospect that our great-grandchildren would live better. Our golden rule need not be, "Jam tomorrow and jam the next day, but never jam today!"

REFERENCES

Barry Bosworth, James S. Duesenberry, and Andrew S. Carron, *Capital Needs in the Seventies*, Washington 1975.

Benjamin Friedman, "Financing the Next Five Years of Fixed Investment," *Sloan Management Review*, Spring 1975, *16*, 51–74.

———, "Discussion" (of **Andrew F. Brimmer** and **Allen Sinai,** "The Effects of Tax Policy on Capital Formation, Corporate Liquidity and the Availability of Funds: A Simulation Study"), *J. Finance*, May 1976, *31*, 309–312

Patric H. Hendershott, "The Impact of a Tax Cut: Crowding Out, Pulling In and the Term Structure of Interest Rates," *J. Finance*, Sept. 1976, *31*.

William E. Simon, Secretary of the Treasury, *Tax Reform*, Public Hearings Before the Committee on Ways and Means, House of Representatives, Ninety-Fourth Congress, First Session, On the Subject of Tax Reform, Part 5, Washington, July 31, 1975.

Allen Sinai, "The Prospects of a Capital Shortage in the U.S.," *Euromoney*, May 1976, as reprinted by Data Resources, Inc., as Economic Studies Series Number 24.

The New York Stock Exchange, *The Capital Needs and Savings Potential of the U.S. Economy, Projections Through 1985*, New York, Sept. 1974.

U.S. Bureau of Economic Analysis, *A Study of Fixed Capital Requirements of the U.S. Business Economy 1971–1980* (prepared under the direction of Beatrice N. Vaccara), Washington, Dec. 1975 (Processed).

Does the United States Save Too Little?

By MARTIN FELDSTEIN*

The division of national income between consumption and saving is probably any economy's most important macroeconomic characteristic. It is significant therefore that the *U.S* saving rate is lower than the rate in almost every other industrial country.[1] While this in itself is neither good nor bad, it arouses concern that the United States may save "too little "

I. Comparing the Reward with the Sacrifice

To know if the United States does save too little we must ask: If we increase our capital accumulation, would the resulting higher level of future consumption compensate sufficiently for the reduced consumption today? The first part of this paper shows how this question can be answered and why I believe the answer is yes Although we usually assume that such questions can be left to individuals, saving decisions are subject to powerful distortions through tax rules and social security. It is necessary to look beyond observed choices and compare explicitly the benfits and costs of additional saving

A. The National Rate of Return on Private Investment

The first step in answering our question is to estimate the rate of return that the *nation* would earn on additional saving, i.e , the effect that foregoing a dollar's worth of consumption would have on the income available for national consumption in the future Although this rate can only be approximated by aggregate statistics, I believe that the available data can provide a useful estimate and an assessment of any significant trends.

In a recent study, Larry Summers and I estimated the potential national rate of return on private investment by the ratio of the pretax capital income (including interest but net of depreciation) to the value of the capital stock in the nonfinancial corporate sector. We were fortunate to have the new Department of Commerce data on annual capital income and capital stock for the postwar period in constant dollars, with the old tax accounting measures of depreciation at historic cost superseded by new measures of depreciation at replacement cost and with inventory profits purged of the distorting effects of inflation

For the entire postwar period 1946–75, we estimated the average net rate of return at 12.4 percent For individual overlapping decades the estimated rates of return were: 1946–55, 13.5 percent; 1950–59, 12 6 percent; 1956–65, 12.4 percent; 1960–69, 13 4 percent; and 1966–75, 11.2 percent.

The lower rate for the most recent decade may suggest a permanent decrease or even the beginning of a secular decline, a possibility stressed in the widely cited paper by William Nordhaus. But such an interpretation ignores the cyclical nature of profits. When Summers and I analyzed the annual rates of return, we found absolutely no evidence of a downward trend in the period through 1969. While the profit rate fell substantially in the following six years, we believe this reflects the abnormal experience of an imported inflation, price controls, and a sharp recession. Indeed, the adverse effect on profits of a low rate of capacity utilization that can be inferred from the experience before 1970 is sufficient to explain the seeming downward trend in profits for the period through 1975. Moreover, the rate of profit

*Professor of Economics, Harvard University The current paper is developed more fully in Feldstein (1976d) I am grateful to the National Science Foundation for financial support

[1]For the twenty-four Organisation for Economic Co-Operation and Development members other than the United States, gross fixed capital formation averaged 24 percent of gross domestic product in the period 1962 through 1973 The *U S* rate was only 17 percent while the rate in Japan was 33 percent The *U S* net national saving ratio of less than 8 percent for this period is also very low by foreign standards.

began to recover in 1975 and will be up again in 1976.

It seems most appropriate to conclude that the national rate of return on private corporate investment is about 12 percent and shows no evidence of a permanent or secular decline.

B. Preference for Present and Future Consumption

I for one would save more if I could obtain a 12 percent return. But as economists we want to say something about the rate at which others would be prepared to substitute future consumption for present consumption. If the amount of future consumption that individuals require to forego present consumption is less than the rate at which investment produces future consumption from current capital investments, we should save more: i.e., *the U.S. saves too little if the rate at which individuals discount future consumption is less than the national rate of return on private investment.*[2]

If there were perfect capital markets and no taxes on capital income, everyone's rate of time discount *(d)* would equal the market rate of interest. With no taxes, this rate would also equal the marginal product of capital. The existing personal and corporate income taxes put a wedge between the national return on capital and the net rate received by savers. As a first approximation, everyone equates his rate of time discount to the net of tax return that he receives. The substantial tax "wedge" makes the consumption discount rate *(d)* substantially less than the pretax national rate of return on additional investment *(r)*. As a simplified example, note that a corporation tax at rate t_c and a personal income tax at rate t_p imply that $d = (1 - t_p)(1 - t_c)r$. Understating the effective tax rates as $t_p = 0.3$ and $t_c = 0.4$ overstates d; still, a national rate of return of 12 percent corresponds to $d = 0.050$. The net return received by investors and therefore their rate of time discount of future consumption is less than half of the corresponding pretax national rate of return.[3]

There is a quite different way to think about comparing consumption at different dates. In cardinalist language, the marginal rate of substitution between consumption at different dates is the ratio of the corresponding marginal utilities of consumption: $MRS_{s,s+1} = U_s/U_{s+1}$ where $U_s = \partial U/\partial C_s$. Ignore for the moment the fact that future consumption is less certain because of the probability of intervening death and psychologically less attractive because of what Pigou referred to as the "faulty telescopic faculty" that causes future pleasures to appear smaller than they are in reality. The marginal utility of consumption nevertheless declines through time because real consumption per capita rises. If, over the relevant horizon, consumption grows exponentially at rate g and the elasticity of marginal utility with respect to consumption is a constant of m, the marginal utility of consumption will also fall exponentially at rate gm. The marginal rate of substitution therefore satisfies $MRS_{s,s+1} = e^{gm}$. Since this derivation ignores both the individual probability of death and the psychological myopia and idealizes the change in consumption as a constant exponential growth, the resulting marginal

[2]Critics of increased saving often ask "Why should we save more to benefit the next generation? They will be richer than we are." This line of argument is quite irrelevant. Although we *can* save more in order to *give* more wealth to the next generation, additional saving can also be purely selfish. We can save more in order to enjoy a higher standard of living in our own retirement or in later preretirement years. At that time we can individually *sell* the capital stock to the next generation and consume its value. If each generation chooses to save more *for its own retirement years*, the capital stock will be permanently higher.

[3]Inflation tends to raise the effective tax on capital income even further and therefore to widen the gap between d and r; see Feldstein, Jerry Green and Eytan Sheshinski. Replacing the mean values that I have been discussing with corresponding "certainty equivalents" yields would lower both r and d but would increase the relative difference because the current absolute difference between the means reflects the portion collected by the government which is pooled and spread and which therefore need not be reduced (or reduced very little) in going from a mean rate to a certainty equivalent rate. In Feldstein (1976d), I discuss the implication of the fact that most individuals do not save by buying corporate stock but by accumulating pension reserves or savings account deposits, the problem of separate borrowing and lending rates, the multiplicity of pretax returns on different investments, and the sensitivity of the return to additional investment.

rate of substitution is best thought of as representing a "planner's time preference" that is appropriate if we wish to ignore the distribution of consumption among individuals including the distribution among individuals of different generations. To distinguish this from the individuals' time preference rate d, I will denote this by δ. thus, $MRS = (1 + \delta) = e^{mg}$. As a quite accurate approximation for the relevant orders of magnitude, δ is mg. Note that δ will be less than d by the annual probability of death and by the discounting of future utility that individuals would later recognize as irrational

A numerical example will help to fix these ideas. Since per capita consumption is growing at about $g = 0.02$, we find $\delta = 0.02m$. If a 10 percent increase in consumption causes its marginal utility to decrease by 20 percent, $m = 2$. The appropriate value of m is clearly a matter of introspective judgment; I think of m as between 0.5 and 1.5 and would find values of m much in excess of 2 to be quite implausible Even with $m = 2$, $\delta = 0.04$ A reasonable adjustment for the probability of death and for Pigovian myopia would still leave d at no more than 0.07 Thus, this direct utilitarian approach, like the analysis of prevailing net of tax asset yields, implies that the rate of time discount of future consumption is probably about half of the social return on additional private investment

II. Why Do We Save Too Little?

I have just discussed how *taxes* on capital income reduce the reward that savers receive for postponing consumption If the government financed the same public spending by a tax that exempted capital income, the reward to savers would rise and the nation's rate of saving would increase.[4] Since eliminating or reducing the tax on capital income would require increasing the effective tax on labor income, the welfare gain from removing the saving distortion would be partly offset by a welfare loss from a greater distortion of labor supply However, detailed calculations indicate that, with plausible but conservative parameter values, the welfare gain would outweigh the loss (Feldstein 1976a). We save "too little" because of taxes in the sense that both saving and economic welfare would increase if the taxes on capital income were reduced and replaced by a tax on consumption with equal yield and equal progressivity

As I explained in previous papers (Feldstein 1974, 1976b,c), *social security* affects saving in two countervailing ways. For someone with fixed retirement plans, social security unambiguously reduces private saving by substituting the promise of future benefits for real private retirement assets. However, social security induces earlier retirement which in itself increases saving Although the net effect is theoretically indeterminate, there is a growing body of econometric studies using both aggregate data and household survey data that shows that social security does substantially reduce private saving and therefore national saving (see Feldstein 1976b).

To understand the nature of the *welfare loss* that occurs when an increase in social security depresses private saving, consider an increase that is small enough to leave unchanged the national rate of return and the rate of time preference Paul Samuelson's model of overlapping generations is a convenient framework for this analysis It shows that if each generation lives for one "year" and aggregate real income grows at rate n, social security "pays" an implicit rate of return of n, i.e , for each one dollar of social security taxes that individuals pay during their "working year" they will receive $1+n$ dollars of benefits in retirement during the "next year" If the one dollar of taxes had instead been invested in real capital accumulation, the return would have been r dollars. The individual thus loses $r - n$ dollars during the "retirement year" per dollar of tax paid (rather than invested) in the previous "working year." The discounted value of that loss at the time that the tax is paid is thus $(r-n)/(1+d)$, where d is

[4]Substituting a consumption tax for our current income tax while keeping the present value of everyone's lifetime tax burden unchanged would be a change in the timing of aggregate tax collections. National saving would increase only if the government adjusted its net surplus to keep real government consumption unchanged A tax on "labor income" (defined to incude the receipt of gifts and bequests) has the same effect on personal consumption but a yet different timing of tax receipts. See Feldstein (1976a)

the individual's rate of time discount.

Consider now a decision to increase social security taxes and benefits by S dollars at time t = 0 and to raise this increment annually at rate n as national income grows. There is an immediate gain of S dollars to the generation of retirees who receive the initial transer without paying any extra tax and a net loss to the present generation of workers and to each future generation; for those who are working in "year" t the value of the net loss is $[(r-n)/(1+d)]S(1+n)^t$. The immediate gain of the current retirees can be compared to the current and future losses by discounting these losses at the time preference rate δ that is appropriate for intergenerational comparisons of consumption. It is easily shown that the present value of the future *net* loss per initial dollar of increase (S) is:[5]

$$(1) \qquad \frac{\text{Loss}}{S} = \frac{r-n}{\delta-n} \cdot \frac{1+\delta}{1+d} - 1.$$

Since $r > n$ and $r > d \geq \delta$, the loss is clearly positive

Readers familiar with Samuelson's analysis may wonder why he reached the very different conclusion that social security would raise the welfare of every generation. Unlike the current analysis, Samuelson's model assumed that no capital goods exist so that real saving and investment is impossible. By extension, whenever r is less than n, the "loss" of each future generation is actually a gain and social security unambiguously raises welfare by reducing real investment (see David Cass and Menahem Yaari). But in the realistic case of $r > n$, the loss depends on the relative magnitudes of r, n, δ and d.[6]

The issue of optimal social security benefits is of course more complex than this simple discussion implies. But the analysis is sufficient to illustrate the basic point: in reducing private saving, social security causes the substitution of a low-yielding implicit intergenerational contract for real capital investment with a higher social yield.

III. Four Wrong Reasons for Saving More

I would now like to contrast the reason I emphasized for saving more—that the benefits greatly exceed the cost—with the main arguments in the recent "capital shortage" debate.[7]

The Capital Gap. The public's attention was drawn to the issue of a "capital shortage" by cries of alarm from the Secretary of the Treasury and others who projected "investment needs" that greatly exceeded forecast saving. Whatever the accuracy of these projections, there is an important sense in which any such forecast of a "capital gap" is misleading. It appears to predict that the demand for capital will continually exceed its supply. Usually when there is excess demand for some good, its price rises until demand and supply are equal. In the capital market, the interest rate and the cost of equity capital should increase until they are high enough to force firms to tailor their aggregate investment demands to the available supply. There will be no "shortfall" of investment funds because the demand for funds will shrink to the available supply.[8]

Full Employment. A quite different notion of a capital shortage is offered by those who believe the capital stock is "too small to provide full employment." Such a view is contrary to both the Keynesian analysis of aggregate de-

[5]This entails the convergence condition that $\delta > n$, an assumption that is not necessarily satisfied Recall that δ may be regarded as the rate of decline of the marginal utility of consumption *per capita* while n is the rate of growth of *aggregate* income. If $\delta < n$, the future losses have an infinitely large present value, limited in reality only by the eventual limit to population growth

[6]If $r = d = \delta$ there would be no present value loss in the case being considered although each generation of workers would lose $(r-n)/(1+d)$ With no tax distortion $(r = d)$ but with $d > \delta$, there would be a loss per dollar of initial tax increase of $(d-\delta)(1+n)/(1+d)(\delta-n) > 0$

[7]There is also the more reasonable but technically false assertion that an increase in the rate of saving is desirable because it causes an increase in the rate of growth of national income Although a higher saving rate does cause a temporary increase in the rate of growth of income, it is better to regard this as a *transition* to a higher level of income. Eventually the rate of growth returns to its original value

[8]Among those who have emphasized the logical contradiction of a "capital gap" are Barry Bosworth, James Duesenberry and Andrew Carron; Benjamin Friedman, Allen Sinai and Roger E Brinner, and Paul Wachtel, Arnold Sametz and Harry Shuford

mand unemployment and the neoclassicial view that the high "permanent" rate of unemployment in the United States reflects adverse incentives that result from government policies and labor market institutions. At best, the notion that more capital can lower the unemployment rate could be rationalized in terms of a temporary situation in which there is no unused excess capacity and no opportunity to use existing capital in a more labor-intensive way; more capital would then be a prerequisite for more employment. Neither of these conditions holds at present: there is substantial evidence of excess capacity and the capital stock can be used in a more labor-intensive way by greater reliance on multiple-shift working

More important, any such argument for more capital to reduce unemployment confuses the occasional desirability of a *temporary increase* in the capital stock with the desirability of a *permanently higher* saving rate and correspondingly large capital stock. In the long run which is the focus of this paper, a larger capital stock would not reduce unemployment because the capital-labor ratio in production would rise as capital became more available.

Price Stability. A similar confusion of temporary increases in investment with permanent increases in the capital stock underlies the price stability case for increased saving. Since some price increases occur when the demand for particular products exceeds capacity output, selective temporary increases in capacity could eliminate this potential source of inflation. But a permanently higher saving rate and a correspondingly higher capital stock would not reduce the frequency or severity of bottlenecks and excess demand. While the larger capital stock would permit a higher level of output, it would also raise the level of wages and capital income and therefore the demand for that output.

International Competitiveness. "A larger capital stock would increase the productivity of U.S. workers. For any given level of real wages, greater productivity means lower prices. And with fixed exchange rates, lower prices mean more exports and fewer imports." This line of reasoning is used to argue that a higher saving rate will improve the long-run balance of trade and, by reducing imports and increasing exports, will "prevent the loss of American jobs to foreign workers."

It should be clear that the argument is faulty at several points. Higher productivity should increase real wages. The level of prices will depend on (among other things) the ratio of the money supply to the level of output and not on productivity or other such "real" variables. Exchange rates do not actually remain unchanged in the long run even when exchange rates are officially "fixed" and certainly vary quite rapidly under the current system of "managed floating" exchange rates. The domestic price level can therefore change without affecting exports and imports. And, finally, if the domestic labor market functions efficiently and aggregate demand is maintained, there will be no relation between the level of net exports and the level of domestic employment. If there is a relation between capital accumulation and export performance, it should be both temporary and weak.

IV. Conclusion

The existence of this session is ample evidence that economists and others are asking whether the United States saves too little. I believe the answer is "Yes": the reward for additional real saving that the nation as a whole would receive would be well worth the current sacrifices.

Debates about the notion of a "capital shortage" have served only to confuse this issue. It is quite appropriate to belive that there is no "capital shortage" but that an increase in the saving rate would be desirable. There are four principal ways in which public policies can increase national saving: government surpluses, change in tax rules, changes in the structure of social security benefits and financing, and reform of the regulation of financial institutions. I hope that future analysis will focus on defining the appropriate mix of these four options.

REFERENCES

Barry Bosworth, James Duesenberry and Andrew Carron, *Capital Needs in the Seventies,* Washington 1975.

David Cass and Menahem Yarri, "Individual Saving, Aggregate Capital Accumulation, and Efficient Growth," in **Karl Shell** (ed.), *Essays in the Theory of Optimal Growth,* Cambridge, Mass. 1967.

Martin Feldstein, "Social Security, Induced Retirement and Aggregate Capital Accumulation," *J. Polit. Econ.,* Sept./Oct. 1974, *82,* 905–26.

———, "Welfare Loss of Capital Income Taxation," *J. Polit. Econ.,* forthcoming, 1976a.

———, "Social Security and Saving: The Extended Life Cycle Theory," *Amer. Econ. Rev.,* May 1976b, *66,* 77–86.

———, "Social Security and Private Savings: International Evidence in an Extended Life Cycle Model" in **Martin Feldstein** and **Robert Inman** (eds.), *The Economics of Public Services,* an International Economic Association Conference volume, forthcoming, 1976c.

———, "National Saving in the United States," in *Investment and Saving for Productivity, Growth and High Employment,* an American Assembly Conference, forthcoming 1976d.

———, **Jerry Green and Eytan Sheshinski,** "Inflation and Taxes in a Growing Economy with Debt and Equity Finance," *J. Polit. Econ.,* forthcoming, 1976.

———, **and Lawrence Summers,** "The Rate of Profit: Falling or Cyclical?", forthcoming, 1976.

Benjamin Friedman, "Financing the Next Five Years of Fixed Investment," *Sloan Management Review,* Spring 1975, *16,* 51–74.

William Nordhaus, "The Falling Share of Profits," *Brookings Papers,* 1974, *1,* 169–208.

Paul A. Samuelson, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *J. Polit. Econ.,* Dec 1958, *66,* 467–82.

Allen Sinai and Roger E. Brinner, *The Capital Shortage,* Lexington, Mass 1975.

Paul Wachtel, Arnold Sametz and Harry Shuford, "Capital Shortages: Myth or Reality," *J Finance,* May 1976.

Some Reflections on Capital Requirements for 1980

By BEATRICE N. VACCARA*

It seems a bit ironic that in the Bicentennial year of this country's history a session of the American Economic Association should concern itself with the problem of capital shortage, for surely this must have been a subject of discussion by colonial economists two hundred years ago. Yet clearly "capital shortages" have not been a chronic problem of the *U.S.* economy, and most of us view this as a problem of more recent vintage. However, some of you may remember that almost twenty-five years ago Wassily Leontief (1953) stressed that the net impact of our foreign trade was to import capital and to export labor. He went on to argue further that because of the high productivity of *U.S.* labor relative to other countries, capital rather than labor was our relatively scarce resource. At the time of Leontief's original work this finding was considered a paradox and many individuals, including myself, were critical of his findings, either on theoretical or empirical grounds. Yet today his contention that labor is our relatively abundant resource and that capital is our scarce resource is regarded by some as an appropriate description of the economic scene.

If I were to interpret today's assignment as an attempt to answer the question—"Will there be a capital shortage?"—my answer would have to be a simple "No," for we are all aware that investment always equals saving and that in a "free market" economy there can be no real shortages of a reproducible good. Of course, I might also want to point out that while, in theory, such a disequilibrium could not persist, at least in the long run, in reality, institutional constraints and distortions do present obstacles to an automatic adjustment to an "optimal" level of capital formation.

My talk today, however, will not discuss such issues but will concentrate on such questions as: 1) What level of fixed nonresidential investment is consistent with our pronounced long-run national objectives of full employment, increasing productivity, environmental cleanup and a drive towards energy conservation and decreasing dependence on foreign sources of petroleum? 2) How does one go about estimating these investment requirements within a consistent *GNP* framework? and 3) How sensitive are the estimates to the various steps of the procedures employed?

The general methodology employed by the Bureau of Economic Analysis (*BEA*) in its capital requirements study was a combination of a macroeconometric forecasting model and a detailed input-output model. Such an approach was used because it was believed that building up the aggregate from detailed industry estimates would not only yield a better total, but would also permit one to "see what was going on," that is, to separate out the various factors that contribute to total capital requirements by business: capital expansion, capital replacement, environmental cleanup and the drive towards energy conservation and self-sufficiency.

The broad steps required for such an approach were as follows: 1) projecting *GNP* and its major components to 1980; 2) translating these aggregate *GNP* projections into detailed industry bills of goods; 3) deriving the Gross Domestic Output requirements by industry associated with this set of final demands (through the use of a projected input-output inverse matrix); 4) estimating the gross capital stock needed to produce the projected industry outputs (by multiplying the projected industry out-

*Associate Director for National Analysis and Projections, Bureau of Economic Analysis, U.S. Department of Commerce.

puts by projected capital/output ratios); 5) estimating the cumulative gross private domestic investment (1971–80) required to assure the necessary 1980 capital stock, allowing for both replacement and expansion of productive capital; 6) estimating the investment requirements related to energy conservation and attempts at energy self-sufficiency; and 7) estimating the additional investment required by each industry to meet existing regulations relating to pollution control and abatement.

Figure 1 provides additional detail on the basic steps used in the estimating procedures. Even a quick glance at this flow chart (which has been simplified somewhat for presentation purposes) indicates that the procedures employed were rather complex and involved substantial data inputs. *BEA* was not alone in providing the data input, but utilized and synthesized information developed elsewhere in the federal government and by private research organizations.[1]

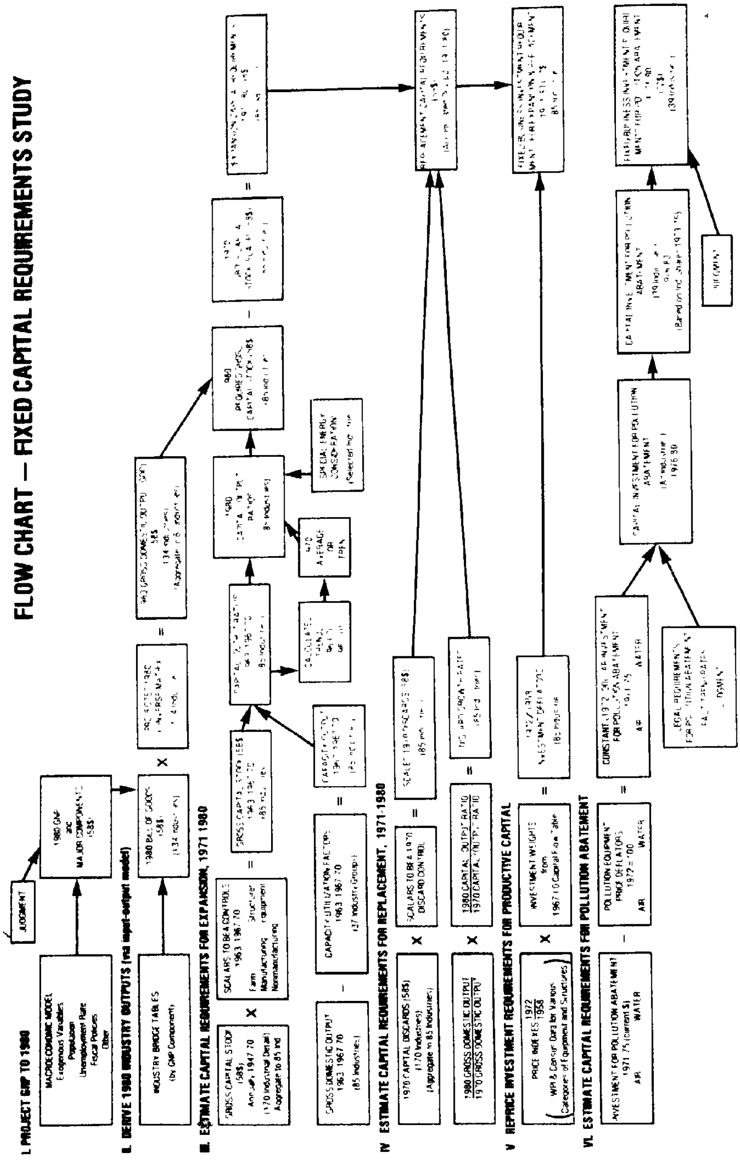
Although the time constraints do not permit me to go into detail on all aspects of the methodology, I would like to highlight one important aspect—the procedure used to drive the 1980 capital/output ratios by industry. The approach was a pragmatic one. In view of the unresolved theoretical debate on the nature of the production function and in the face of inconclusive and sometimes conflicting empirical evidence regarding whether or not technical progress is neutral, capital augmenting, or labor augmenting, this seemed the wisest course.

The first step of the procedure for projecting capital/output ratios involved the calculation of historical time series on industry capital/output ratios. In constructing the historical ratios, two important elements had to be considered: The output measures had to reflect input-output concepts and definitions (since these ratios were to be applied to output totals derived through the use of an input-output inverse matrix) and second, adjustment had to be made to allow for the degree of utilization of the existing capital stock. If such utilization adjustments were not made, for many industries the resulting ratios could overstate the size of the capital stock required to produce a given volume of output.[2]

Because of data limitations, capital/output ratios for the entire postwar period could be developed only for the private economy as a whole; industry ratios could be developed for only the selected years, 1963, 1967, 1968, 1969, 1970. The ratios for the total private economy indicated a mixed picture as regards trends during the postwar period; moreover, this picture differed, depending upon whether or not capital was adjusted for underutilization. During the period 1947–61, there was a clear-cut downward trend in the adjusted ratios, while the period 1962–69 showed a reversal in the direction of this trend, with the 1973–74 ratios about equal to the 1969 ratio. However, given the impact of shifting industrial mix (due to both cyclical and more long-run factors) on the observed overall capital/output ratio for the total private economy, it was deemed inadvisable to assume the absence of clear-cut trends for individual industries. Hence, the capital/output ratios for 1963 and for the years

[1]Because of the necessity of completing the capital requirements study within a short time span, *BEA* did not make independent projections of the *GNP* components, the industry bills of goods and the associated industry output levels, but utilized "revised" 1980 output projections by industry developed by the Bureau of Labor Statistics (*BLS*). (See *BLS*, 1975 and 1976.) The historical data (1947–70) on captial stocks by detailed input-output industry were developed by Jack Faucett Associates. Information on production and capital/output ratios in the energy-producing industries was obtained from the Project Independence Report of the Federal Energy Administration.

[2]Deflated outputs were adjusted for capacity utilization (*CU*) before computation of the capital/output ratio. *CU* rates developed by the Wharton School were used for nonmanufacturing industries and *BEA* rates were used for manufacturing industries—actual utilization rates and "actual utilization rates as a percent of preferred utilization rates." The latter concept was used for purposes of this study. These rates, which are somewhat higher than acutal utilization rates, allow for the fact that not all existing capacity is economically efficient. These rates were deemed conceptually more consistent with the Wharton rates which are derived via a procedure which equates capacity with peak output levels.


FLOW CHART — FIXED CAPITAL REQUIREMENTS STUDY
I. PROJECT GNP TO 1980
JUDGMENT
MACROECONOMIC MODEL
Exogenous Variables
Population
Unemployment Rate
Fiscal Policies
Other
1980 GNP
and
MAJOR COMPONENTS
(58$)
II. DERIVE 1980 INDUSTRY OUTPUTS (via input-output model)
INDUSTRY BRIDGE TABLES
(by GNP Component)
1980 BILL OF GOODS
(58$)
(134 Industries)
X
=
III. ESTIMATE CAPITAL REQUIREMENTS FOR EXPANSION, 1971 1980
GROSS CAPITAL STOCK
(58$)
Annually 1947 70
(170 Industrial Detail)
Aggregate to 85 Ind
X
SCALARS TO BEA CONTROLS
1963 1967 70
Farm
Manufacturing
Nonmanufacturing
Structures
Equipment
GROSS DOMESTIC OUTPUT
1963 1967 70
(85 Industries)
CAPACITY UTILIZATION FACTORS
1963 1967 70
(37 Industry Groups)
AVERAGE
OR
TREN
SPECIAL ENERGY
CONSERVATION
(Selected Industries)
IV ESTIMATE CAPITAL REQUIREMENTS FOR REPLACEMENT, 1971-1980
1970 CAPITAL DISCARDS (58$)
(170 Industries)
(Aggregate to 85 Industries)
SCALARS TO BEA 1970
DISCARD CONTROL
1980 GROSS DOMESTIC OUTPUT
1970 GROSS DOMESTIC OUTPUT
1980 CAPITAL OUTPUT RATIO
1970 CAPITAL OUTPUT RATIO
V REPRICE INVESTMENT REQUIREMENTS FOR PRODUCTIVE CAPITAL
PRICE INDEXES 1972/1958
WPI & Census Data for Various Categories of Equipment and Structures
INVESTMENT WEIGHTS
from
1967 I-O Capital Flow Table
VI. ESTIMATE CAPITAL REQUIREMENTS FOR POLLUTION ABATEMENT
INVESTMENT FOR POLLUTION ABATEMENT
1971 75 (current $)
AIR WATER
POLLUTION EQUIPMENT
PRICE DEFLATORS
1972 = 100
AIR WATER
CAPITAL INVESTMENT FOR POLLUTION
ABATEMENT
1976 80
JUDGMENT

1967–70 for each of 85 industries were examined to determine if there were any clear-cut trends; where such trends were evident, a continuation of these trends to 1980 was assumed. For other industries, the 1970 ratio or an average of the ratios for the 1967–70 period was used for 1980. It was found that 16 industries, accounting for 13 percent of the 1970 total capital stock, showed upward trends, while 3 industries, accounting for 7-½ percent of the 1970 total capital stock, showed declining trends. The remaining 66 industries evidenced no clear-cut trends in the capital/output ratios for the recent historical period.

Although there are many other aspects of the methodology critical to the findings and hence worth examining, given the fact that the methodology is fully detailed in the *BEA* report (Vaccara 1975), I will turn next to an examination of the results. The major findings of the study were as follows:

1) In order to assure a 1980 capital stock sufficient to provide for increasing productivity, full employment levels of output, pollution abatement and decreasing dependence on foreign sources of petroleum, during the period 1971–80, nonresidential fixed investment (in 1972 prices) needs to total $1,473 billion or 11 4 percent of *GNP*.

This percentage is higher than the 10.4 percent characteristic of the 1965–70 period and the 1971–74 period [3] Moreover, given actual cumulative investment during 1971–74, for the remaining six years, 1975–80, fixed capital investment must total $987 billion or 12.0 percent of cumulative *GNP*. Of this nearly one trillion dollars of additional investment requirements, roughly 52 percent is for replacement, 45 percent is for expansion of capacity and 3 percent is for pollution abatement.

[3]The *BEA* study of capital requirements was undertaken in the summer and fall of 1975; at that time the benchmark revision of the national accounts had not been completed Thus, the numerical values in the orginal report, some of which are repeated here, do not reflect these revisions Based on the revised data, the 1971–74 ratio of nonresidential fixed investment to *GNP* was 10.2 percent rather than 10 4 percent

2) These requirements for an increased share of *GNP* devoted to investment are associated with three factors:

a) Changing technology in selected industries such as agriculture, iron ore mining, nonferrous metals manufacturing, communications equipment, business services and auto repair services, where the capital/output ratios have been increasing, adds $118 billion to the cumulative 1971–80 constant dollar investment requirements This is only partially offset by industries with declining trends in capital/output ratios and thus technological change is a major cause of the need for an increased share of *GNP* devoted to fixed investment adding $82 billion (net) to the cumulative investment total as compared to a fixed 1970 technology.

b) Increased needs for capital investment in the petroleum mining, electric utility, and other industries as a result of the "energy situation" adds about $58 billion to the cumulative 1971–80 investment total.

c) Investment in pollution abatement equipment as a consequence of legislation relating to "clean air" and "clean water" adds about $50 billion to the total.

In addition to the three factors mentioned above, all of which worked to increase the capital/output ratio for the aggregate economy, a fourth factor—the shifting industrial mix of *GNP*—worked to reduce the aggregate capital requirements per dollar of *GNP*. This reduction resulted primarily from a decline in the relative importance of output demanded from the highly capital intensive sectors such as petroleum and natural gas mining and transportation. Thus, if the industrial mix of goods and services which consumers, business and government demanded remained unchanged between 1970 and 1980, the cumulative investment requirements would be even greater—by an additional $75 billion.

If it were not for the factors mentioned, that is, if the productive technology and the mix of

goods and services demanded remained unchanged between 1970 and 1980, if pollution control and abatement legislation had not been passed, and if we were content to continually increase our dependence on foreign sources of petroleum, capital requirements in the decade of the 1970's would average only 10.5 percent of *GNP* or about $1,360 billion in 1972 prices

Since the completion of the original capital requirements study last fall, data for 1975 have become available. We also have some fairly good indications of the likely investment path for 1976. These figures indicate that fixed nonresidential investment is falling behind the target rate. In 1975 fixed nonresidential investment (in 1972 dollars) totaled $111 4 billion or 9.6 percent of constant dollar *GNP* and it is estimated that in 1976, investment will be a similar percentage of constant dollar *GNP* Thus, these figures indicate that even more investment will be required in the remaining years of this decade if the goals of full employment, increasing productivity, pollution abatement and less dependence of foreign energy sources are to be realized.

The results just described are sensitive to the methodology employed—to the assumed level of full employment *GNP*, the industry mix of final demand and the trend rate used to project the capital/output ratios The estimates are particularly sensitive to the industry mix of *GNP* For example, direct and indirect capital requirements per dollar of output of the railroad industry are almost nine times that of the leather and footwear industry

The projected estimates of capital requirements may be on the high side. There are several reasons for this belief.

1) The 1980 "full employment" *GNP* of $1,575 billion (in 1972 prices) projected by the *BLS* is about 3–4 percent higher than an alternate estimate recently developed by *BEA*. (This is due in part to *BEA*'s use of a higher unemployment rate for "full employment"—5.2 percent as compared to 4.7 percent—and to a somewhat lower implied productivity growth rate.)

2) The historical capital/output ratios in some nonmanufacturing industries may be overstated because the capital stock was not adjusted for underutilization due to the lack of available data on capacity utilization.

3) The use of the period 1963–70 for determining trend rates in the industry capital/output ratios may overstate the upward trends during 1970–80 because interest rates were considerably lower during that period than the more recent period. Thus, the observed tendencies in some industries for "capital deepening" may not continue at the same average annual rate.

It should be noted, however, that there are some elements of our methodology which could cause an understatement of the derived capital requirements and which could thus offset some of the factors leading to an upward bias. One possible source of a downward bias is our inability to evaluate the impact of higher energy prices on the technology of industries which are heavy consumers of energy. For these industries we may have understated the rate of replacement of investment as companies discard energy intensive machinery and facilities in favor of those which are more energy efficient. Additional downward bias may be due to our failure to allow for any additional capital requirements associated with the Occupational Safety and Health Act. No data are presently available that could provide a basis for making reliable projections in this area Finally, some downward bias may also be caused by our implicit assumption that companies operating below their preferred utilization rate will increase output by increasing the rate of utilization of their existing capital stock (up to the preferred rate) before expanding their capital stock

Despite the deficiencies just described, I consider the estimates to be relatively accurate and believe the message they convey. This message indicates that the goals which we have set for ourselves: full employment by 1980, increasing productivity, environmental clean-up and decreasing dependence on foreign sources of energy cannot be achieved unless we devote more

of the pie to investment than we have devoted during the past decade. If we are not prepared to do this, and some question whether or not we should, we have to be prepared to modify some of our objectives.

The most likely consequence of a failure to achieve the desired rates of capital investment would be to delay the achievement of these goals to some time beyond 1980, and to force a modification of some of our stated goals. This seems to be particularly true of our objective to decrease our dependence on foreign sources of petroleum. Should we attempt to achieve all these objectives by 1980 without an extra stimulus to investment, it is highly likely that capacity pressures and even bottlenecks could occur. However, these capacity pressures will not necessarily occur in the paper, primary metals, and chemical industries as was the case in the late 1973 peak period when all of these industries were utilizing 95 percent or more of their preferred capacity. Since that time, these industries have had substantial increases in investment—investing at a rate much higher than the overall average for all industries. Although it is difficult to evaluate how much the additional investment in 1974 and 1975 increased productive capacity, since a sizeable share undoubtedly went for replacement investment and for capital expenditures for pollution abatement and control, there are indications that capacity has expanded. For example, information for the chemical industry indicates that despite the fact that production levels in March 1976 were almost 5 percent above those of December 1973, the March 1976 capacity utilization rate was .89 as compared to .96 in December 1973.

One final word of caution, although my analysis indicates that appropriate government policy over the next several years should aim at increasing the business investment share of *GNP*—this is not necessarily the appropriate policy for the longer run. For much of this need for additional capital formation is a one-time thing—that is, once the pollution abatement equipment is in place, once the alternative energy consumption and production technologies have been developed, capital formation can return to its more historic share.

REFERENCES

Marie Hertzberg, Alfred Jacobs and Jon Trevathan, "The Utilization of Manufacturing Capacity, 1965–1973," *Survey of Current Business*, July 1974.

Ronald Kutscher, "Revised BLS Projections to 1980 and 1985: an overview," *Monthly Labor Rev.*, March 1976, *99*, 3–8.

Wassily Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proceedings* of the American Philosophical Society, Sept. 1953.

———, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov 1956, *38*, 386–407. (Also see *Rev. Econ. Statist.*, Supplement Feb. 1958, for comments by Stefan Valavanis, Romney Robinson, George Elliott and Beatrice Vaccara.)

Beatrice Vaccara, *A Study of Fixed Capital Requirements of the U.S. Business Economy, 1971–1980*, Bureau of Economic Analysis, Washington, Dec. 1975.

Federal Energy Administration, *Project Independence Report*, Nov. 1974.

Jack Faucett Associates, *Methodology for Constructing Gross and Net Capital Stock Series for Input-Output Sectors*, Sept 1967.

U.S. Bureau of Economic Analysis, *Fixed Nonresidential Business Capital in the United States, 1925–1973*, Washington, Jan. 1974.

U.S. Bureau of Labor Statistics, *The Structure of the U.S. Economy in 1980 and 1985*, Bulletin 1831, Washington 1975.

ANALYSIS OF DOMESTIC INFLATION

The Theory of Domestic Inflation

By ROBERT J. GORDON*

Authors and readers of the thousands of articles and books published on inflation during the past decade may regard as audacious any attempt to survey the theory of domestic inflation in 3,000 words. But far from requiring an apology, this format forces concentration on central issues and justifies skipping second-order questions. More leisurely expositions and extensive bibliographies are provided in recent surveys by David Laidler and Michael Parkin and by Robert J. Gordon (1976). The ground rules for this paper are a limitation to theory rather than empirical tests, to closed rather than open economies, and to causes of inflation rather than costs, consequences, or cures.

I. Inflation and Money in the Long Run

A simple set of definitions helps to separate noncontroversial from controversial issues. We begin with a national income identity, expressed in growth-rate form.

$$y \equiv p + q, \tag{1}$$

where lower-case letters represent rates of growth, and y, p, and q stand for, respectively, the rates of growth of nominal income, the aggregate price deflator, and real output. Subtracting the long-term trend growth rate of capacity (q^*) from both sides of (1), we obtain.

$$y - q^* \equiv p + q - q^*, \tag{2}$$

or

$$\hat{y} \equiv p + \hat{q},$$

where $\hat{y} = y - q^*$, and $\hat{q} = q - q^*$. Arthur Okun (1962) was the first to establish the statistical relation now widely known as "Okun's Law" between the current unemployment rate (U), last period's unemployment rate (U_{-1}), and the output growth deviation ($\hat{q}$):

$$U = U_{-1} - \hat{q}/a, \tag{3}$$

where a is a constant, roughly equal to 3.0 in the United States. When (3) is solved for $\hat{q}$, the result is substituted into (2), and then (2) is solved for the rate of inflation (p), we have:

$$p = \hat{y} + a\,(U - U_{-1}). \tag{4}$$

The sources of change in y can be decomposed if we once again invoke an identity:

$$\hat{y} \equiv \hat{m} + v, \tag{5}$$

where $\hat{m}$ is the growth rate of money adjusted for capacity growth ($\hat{m} = m - q^*$), and v is the growth rate of velocity. Combining (4) and (5), we obtain:

$$p = \hat{m} + v + a\,(U - U_{-1}) \tag{6}$$

Once the economy has settled down at any given unemployment rate ($U = U_{-1}$), the rate of inflation depends only on the adjusted growth rate of money ($\hat{m}$) and the growth rate of velocity (v). Shifts in fiscal policy can cause one-time-only changes in velocity, as even Milton Friedman (1966b) recognized long ago, but cannot cause permanent changes in the growth rate of velocity. Innovations in transactions technology, as well as an income elasticity of the demand for money differing from unity, could make v positive or negative, but these factors appear to exhibit only modest changes

*Professor of Economics, Northwestern University

insufficient to account for marked accelerations or decelerations in inflation.

Changes in the adjusted growth rate of money are thus isolated as a necessary concomitant of long-run changes in the inflation rate. It is in this carefully qualified sense that Friedman (1966a, p. 18) correctly labelled inflation as "always and everywhere a monetary phenomenon." But despite the attempts of some less subtle monetarists to treat this quotation as settling all questions, in fact it represents only a starting point. Accelerations in monetary growth are not usually autonomous whims of central bankers. In most classic wartime or postwar money-fueled inflations and hyperinflations, the role of the monetary authority has been passively to finance deficits resulting from the unwillingness or inability of politicians to finance expenditures through conventional taxation. In the same way, a "cost push" by unions or firms must be ratified continuously by the monetary authority if inflation is to continue

A more general view, explicitly set out in Melvin Reder's classic analysis, attributes inflation to the passivity of the monetary authority in the face of a "tripartite" set of pressures emanating from all groups in society—labor, management, and government R. J. Gordon (1975c) extends this theme by distinguishing the "demand for inflation," i e., monetary accomodation, caused by government's refusal to tax and by pressure groups which attempt to increase their income share, from the "supply of inflation," the degree of response to these pressures, a result of the political balancing of the votes likely to be lost from higher inflation, as against the vote cost of the higher unemployment consequent upon a policy of nonaccomodation.

II. The "Missing Equation"

For anything other than long-run analysis, equation (6) is incomplete. Even if $\hat{m}$ and v are known, there are two remaining unknowns (p and U) but only one equation. A decade ago it was usual to close the model by adding a Phillips curve:

$$p = bp^e + f(U),\ 0<b<1, f'<0. \tag{7}$$

Together (6) and (7) determine a menu of p,U combinations for different $\hat{m}$. It was common in the United States for economic advisers to Democratic Presidents to recommend a combination with higher p and lower U than the target of Republican advisers.

Friedman (1966a, p. 60) was the first explicitly to reject (7) and to state that "there is no long-run, stable trade-off between inflation and unemployment." On the grounds that workers supply labor by evaluating the expected real value of a wage offer, and that the expected and actual price levels cannot diverge in equilibrium, Friedman (1968) and Edmund Phelps argued that in equilibrium with $p = p^e$ only a single "natural rate of unemployment" (U^N) is possible:

$$p = p^e + g(U - U^N),\quad g'<0,\ g(0)=0. \tag{8}$$

The "natural rate hypothesis" (*NRH*) as embodied in (8) completely changed the framework of stabilization policy. No longer could an Adminstration choose its own favorite point on the p,U tradeoff curve. A rate of unemployment below U^N could not be achieved by aggregate demand policy through manipulations of $\hat{m}$, because inflation would continuously accelerate as long as p^e responds to past changes in p:

$$p^e = h(p_{-1}, p_{-2}, \ldots\ ). \tag{9}$$

A permanent reduction in actual unemployment could be achieved without accelerating inflation only by operating directly on U^N, through manpower programs and other subsidies to reduce worker-job mismatch, and through reductions in the minimum wage and in other barriers to the flexibility of relative wages. It was not widely understood that the *NRH* did not establish a link between inflation and money where none existed before. Instead, p and $\hat{m}$ are linked together in (6), whether or not the "missing equation" is provided by the old-fashioned tradeoff curve

(7) or the *NRH* (8).

III. Short-run Price Inflexibility and the Role of Contracts

An important criticism of the *NRH* has been its apparent lack of validation in recession and depression episodes.

Combining (8) and (9), a deceleration of inflation requires that actual U exceed U^N, since p^e cannot fall until p itself first experiences a decline. A period during which U remains above U^N for a substantial period should be characterized by an accelerating decline in p. But during the Great Depression the unemployment rate remained above 8.5 percent for twelve straight years in the United States without the slightest sign of such an acceleration. If the function $g(\ )$ in (8) were completely flat for high values of U, the *NRH* would remain valid only as long as U were kept below the flat range. Even if $g(\ )$ retains its negative slope in the range of U relevant for current policy, a relatively gentle slope nevertheless would make extremely costly any attempt to "beat the inflation out of the system" by the deliberate creation of a recession.

Until recently the apparent downward inflexibility of prices during periods of high unemployment constituted an empirical phenomenon in search of a theory. Okun (1975) distinguishes between "auction" markets (wheat, peso futures) with instantaneous market clearing and "customer" markets in which economic incentives induce long-term contractual arrangements, infrequent price changes, and quantity rationing. Costly search makes customers willing to pay a premium to do business with customary suppliers. Firms, in turn, have an incentive to maintain stable prices to encourage customers to return, using yesterday's experience as a guide. "A kind of intertemporal comparison shopping" discourages firms from raising price in response to short-run increases in demand or decreases in productivity in order to avoid giving customers an incentive to begin exploring. Prices are not completely sticky, however. Widespread knowledge shared by customers and firms that costs have increased permanently allows price increases without providing an incentive for search, as was evident in the rapid response of final goods prices to the energy cost explosion of 1974.

While R. J. Gordon's (1975b) results support at least some role for changes in demand, nevertheless Okun's basic message is validated by the overwhelming share of the total variance of aggregate price inflation which is explained by changes in "standard" unit labor cost (defined for trend rather than actual productivity). Thus the search for an adequate theory of the downward inflexibility or inertia of inflation in the face of deep recessions and depressions turns to the labor market. Substantial attention has been attracted by the theory of implicit labor contracts independently developed by Costas Azariadis, Martin Baily, and Donald Gordon. Firms and workers engage in long-term contractual arrangements, which may be implicit and unwritten, and which specify wage rates in advance. Entrepreneurs are self-selected individuals who are relatively indifferent towards risk and are willing to provide insurance services for their risk-averse employees in the form of a fixed wage rate.

At present the wage contract models are incomplete and subject to criticism. R. J. Gordon (1976, p. 209) pointed out that the Azariadis-Baily-D. Gordon theory could not explain fixed-wage contracts without relying on government transfer payments paid to workers during unemployment, thus providing them with a higher total income over the cycle than they would receive if the wage varied to clear the labor market continuously. But government transfers would induce firms to respond to a recession in demand by laying off workers rather than cutting their wages even without any contractual arrangements, making the contract idea itself irrelevant. Robert Barro also makes the important point that the adoption of fixed-wage contracts imposes dead-weight losses on participants by creating a divergence between the marginal product of labor and the marginal value of time. It is to the advantage of both

firms and workers to maintain employment at its market-clearing level to maximize the total available product pie.

Ongoing theoretical work attempts to "rescue" the fixed-wage contract from these and other criticisms. Herschel Grossman has analyzed the attempt by firms to minimize the "default risk" of workers jumping from the fixed-wage labor contract into the auction part of the labor market when demand is high. Fruitful ideas introduced by various authors include the preference by firms for the relative certainty of the cost reduction achieved by layoffs compared to the uncertainty of the worker's response to a wage cut, and perhaps most important, the role of employer profits made on the specific human capital of experienced employees, leading firms to maintain the wage rate of experienced employees, while achieving lower costs in a recession by laying off the least profitable inexperienced employees. The consensus appears to be shifting toward worker heterogeneity in the form of differential risk of default, and differential endowments of specific human capital, as the most important elements motivating sticky wages, layoffs, and implicit contracts, and away from the completely homogeneous risk-averse workers featured in the earlier Azariadis-Baily-D Gordon approach.

Whatever the precise details of the theory which explains wage and price inflexibility, the implications of such stickiness have been worked out in great detail by Barro and Grossman. Starting from an initial level of output (Q_0) and prices (P_0), let a decline in aggregate demand cut the "market clearing" price level (P^*) at which Q_0 would be purchased. If the price level remains at P_0, firms want to produce as much as before but face a constraint on the amount which can be sold. Even if P drops below P_0, there will still be a sales constraint as long as P remains above P^*. In the labor market the sales constraint forces firms to hire fewer workers than they would prefer at today's too-high sticky wages and prices. The requirement for the sales constraint to be lifted, and for firms to resume operating on their voluntary output supply and labor demand schedules, are (a) an increase in aggregate demand which raises P^* back up to P, or (b) the passage of enough time to allow P to sink down to equal P^*.

IV. The Challenge of Rational Expectations

The Application of Rational Expectations to Economic Policy (*AREEP*) constitutes a radical contribution to the theory of the *short-run* determinants of unemployment and inflation. The *AREEP* model begins with (6) above, often assuming $v = 0$ to simplify the exposition, and thus has no bearing on our previous analysis of the *long-run* connection between p and $\hat{m}$. Equation (6) is combined with the "Lucas supply function" (see Robert Lucas), which limits the source of output and unemployment changes to purely voluntary responses of firms and workers to deviations between actual and expected inflation:[1]

$$U = U^N + g^{-1}(p - p^e). \tag{10}$$

The supply function (10) is simply an inverted version of (8), describes the same long-run equilibrium conditions, and is implicit in expositions of the *NRH* by Friedman (1968) and others. While the idea of rational expectations has been fruitfully applied to the behavior of financial, primary commodity, and other "auction" markets, we argue here that *AREEP* goes badly astray by using (10) as a description of the conditions necessary for short-run output changes in the portion of the economy dominated by "customer" or "contract" markets and sluggish price adjustment.

Expectations are rational when the expectational error ($p - p^e$) is unrelated to all information (I_{-1}) available when expectations were formed, including the autoregressive structure of all variables. The information set I_{-1} includes (6), which (when $v = 0$ and U is constant) implies:

$$p = \hat{m}, \text{ and } \quad p^e = \hat{m}^e. \tag{11}$$

[1]It is customary to include stochastic error terms in the structural equations (6) and (10), but no essential conclusions are changed by omitting these terms in this exposition

Substituting (11) into (10), we have:

$$(12) \qquad U = U^N + g^{-1}(\hat{m} - \hat{m}^e)$$

Thus the monetary authority cannot influence unemployment, even in the short run, unless it acts in an unpredictable way. If it simply responds to an event by a formula known to the public in the previous period as part of the information set I_{-1}, the public will shift its expectation $\hat{m}^e$ by the exact amount of the change in $\hat{m}$, the difference $(\hat{m} - \hat{m}^e)$ will be zero, and unemployment will not change.

The preceding argument has received widespread attention since its formalization by Thomas Sargent and Neil Wallace. It requires for its validity that the price level (P) respond instantaneously to any change in the market-clearing price (P^*), as occurs in (11). When P is sticky and fails to drop instantly to P^*, the firm faces a sales constraint and cannot operate along its voluntary Lucas supply curve (10). Price inflexibility rules out the supply curve and with it the expositions of *AREEP*, all of which to date are built on it. The *U.S.* evidence in favor of sluggish price adjustment is strong. Two of the many studies include R. J. Gordon's (1975b) reduced-form regression between p and past values of m in the postwar United States, which has a mean lag of four years.[2] And Robert Hall has shown that only two percent of the quarterly variation in United States unemployment during 1954–74 remains unexplained in a simple two-quarter autoregression, in contrast to (10) above, in which U can differ from U^N only by the serially uncorrelated random error $(p - p^e)$.

Bennett McCallum has tried to argue that "recognition of price level stickiness does not, in and of itself, negate the Lucas-Sargent Proposition." His argument and its defects are most transparent for the extreme case of completely rigid prices in which $p = 0$ and a rational expectation $p^e = 0$ as well. The expectation error $(p - p^e)$ in (10) is zero, and thus unemployment is unaffected by any aggregate demand policy. But consider a policy which cuts nominal expenditure by half from E_0 to $.5E_0$. According to the McCallum argument, if prices are rigid the price level (P), unemployment, and output (Q) remain at their original level. If originally $E_0 = P_0Q_0$, now $E_1 = .5P_0Q_0$. Production is double the level of sales, and so an involuntary accumulation of inventories occurs and continues as long as E remains low and P remains rigid. Retention of the Lucas supply function in the face of price rigidity thus leads to the counterfactual conclusion that businessmen never cut production in response to involuntary inventory accumulation!

There is nothing wrong with the assumption of rational expectations itself, nor with its fruitful application to financial markets. But in light of widespread evidence that, except in a few scattered auction markets, prices adjust sluggishly to the market-clearing level in response to demand and supply shocks, it is hard to avoid the conclusion that for short-run analysis the Lucas supply function and with it *AREEP* should be relegated to the same scrap heap of discarded ideas where lie the earlier classical models of perfect market clearing laid to rest by Keynes forty years ago.

V. Cost Push, Controls, and Supply Shocks

Much attention in the popular press has been devoted to the positive correlation of inflation and unemployment during some years of the 1970's, and the alleged failure of economists to explain it. The straw man being attacked has only one arm, equation (8) of our two-equation inflation model, and lacks its other arm, equation (4). Further, inflation is necessarily negatively correlated with unemployment in (8) only when p^e is fixed. Inflation can increase while

[2] Some *AREEP* theorists have pointed out another interpretation of my equation, that it represents a relation between p and m^e, with the lag distribution on m representing the adaptive formation of the expectation m^e. It is true that the long lag might represent expectation formation, not sluggish price adjustment. But then why should expectations on money take many years longer to form than expectations on inflation itself, which in interest rate regressions appears to be described by a mean lag of one year or less?

unemployment is rising, as in 1970 and early 1971, if expectations are formed adaptively and p^e is still rising in response to past realizations of p.

In contrast to equation (8), the dynamic supply schedule which plots a negative relation between p and U for given p^e and U^N, equation (4) is a dynamic demand schedule which plots a *positive* relation between p and U for given $\hat{y}$ and U_{-1}. Any event which shifts the supply curve up a fixed demand curve raises p and U simultaneously. We introduce the shift factor (Z) explicitly into (8):

$$(13) \qquad p = p^e + g\,(U - U^N) + Z.$$

Z might be a cost-push pressure by unions, oil sheiks, or bauxite barons. As long as the authorities hold $\hat{y}$ constant, inflation and unemployment will increase simultaneously. The imposition of price controls may introduce a negative value of Z, which with $\hat{y}$ constant will cause inflation and unemployment to decrease simultaneously, as in the pre-election boom of 1971–72. The termination of controls raised inflation and unemployment simultaneously in 1974 R. J. Gordon (1975a) has shown in this context that crop failures or other supply shocks in general have multiplier effects which spread the loss of output into the nonfarm sector.

VI. Inertia and Policy Options

The same downward inertia of price adjustment which vitiates the conclusions of *AREEP* poses obstacles for policymakers. An economy inheriting a substantial fully anticipated inflation and operating at the natural unemployment rate has two problems—how to achieve price stability and how to reduce U^N to allow the creation of jobs for disadvantaged groups suffering from high unemployment rates. The direct remedy for inflation is the creation of a recession, which reduces p below p^e and allows the adaptive expectation of p^e to drift downwards. The permanent benefits of lower inflation must be weighed not only against the transitory output costs of a recession which might last for years, but against the permanent wealth loss caused by the recession-induced drop in saving.

Another remedy is the direct control of wages and prices. Price controls by themselves misallocate resources without permanently reducing inflation, because prices tend to be tied so closely to wage costs. Wage controls by themselves have proven to be politically infeasible; the present British experiment is possible only because it is structured to achieve a massive redistribution of income away from the rich. Recent proposals to "sell" wage controls include clever tax schemes designed to offset the inevitable short-term losses of real income of workers who agree to allow their wages to be controlled.

Finally, the ongoing inflation can be accepted rather than resisted by allowing for the full indexing of financial assets, labor and product contracts, and all nominal dollar amounts (tax brackets, maxima, minima) written into private and government regulations Preliminary research by Joanna Gray and others indicates that full indexing increases macroeconomic stability if the economy only suffers from demand shocks, but in the presence of supply shocks aggravates both inflation and recession. Thus from a social standpoint full indexing is not optimal, but as yet economists have failed to explain why private institutions have provided such an incomplete menu of indexed assets, liabilities, and contracts.

REFERENCES

Costas Azariadis, "Implicit Contracts and Underemployment Equilibria," *J. Polit. Econ.*, Dec. 1975, *83*, 1183–1202.

Martin N. Baily, "Wages and Employment Under Uncertain Demand," *Rev. Econ. Stud.*, Jan. 1974, *41*, 37–50.

Robert J. Barro, "On Long-Term Contracting and the Phillips Curve," University of Rochester, unpublished, Dec. 1975.

——— **and H. Grossman,** *Money, Employment, and Inflation*, Cambridge, 1976.

Milton Friedman, "What Price Guideposts?" in G. P. Shultz and R. Z. Aliber, eds., *Guidelines: Informal Controls and the Market Place* (Chicago, 1966a), 17–39 and 55–61.

———, "Interest Rates and the Demand for Money," *J. Law Econ.*, Oct. 1966b, 9.

———, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

Donald F. Gordon, "A Neo-classical Theory of Keynesian Unemployment," *Econ. Inquiry*, Dec. 1974, *12*, 431–59.

Robert J. Gordon, "Alternative Responses of Policy to External Supply Shocks," *Brookings Papers*, 1975a, *6*, 183–206

———, "The Effect of Aggregate Demand on Prices," *Brookings Papers*, 1975b, *6*, 613–62.

———, "The Demand for and Supply of Inflation," *J. Law Econ.*, Dec 1975c, *18*, 807–36

———, "Recent Developments in the Theory of Inflation and Unemployment," *J Mon. Econ.*, Apr. 1976, *2*, 185–219

Herschel Grossman, "Risk Shifting and Reliability in Labor Markets," *Scand. J. Econ*, forthcoming, 1977

Joanna Gray, "Wage Indexation: A Macroeconomic Approach," *J. Mon Econ.*, Apr. 1976, *2*, 221–36.

Robert E. Hall, "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers*, 1975, *6*, 301–35.

David Laidler and J. Michael Parkin, "Inflation: A Survey," *Econ. J.*, Dec. 1975, *85*, 741–809.

Robert E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

Bennet McCallum, "Price Level Stickiness and the Feasibility of Monetary Stabilization Policy with Rational Expectations," *J. Polit. Econ.*, forthcoming.

Arthur Okun, "Potential GNP: Its Measurement and Significance," *Proc. Amer. Statist. Assn.*, 1962, 98–116

———, "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers*, 1975, *6*, 351–90.

Edmund S. Phelps, "Phillips Curves, Expectations of Inflation, and Optimal Unemployment Over Time," *Economica*, August 1967, *34*, 254–81.

Melvin W. Reder, "The Theoretical Problems of a National Wage-Price Policy," *Can. J. Econ.*, Feb 1948, 46–61

Thomas J. Sargent and Neil Wallace, "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr 1975, *83*, 241–57.

Measuring Prices—and Wages

By JACK E. TRIPLETT*

One reason economists have had so much difficulty explaining and predicting inflation is conceptual inadequacies and measurement errors in the variables used in the analysis. Possible errors, biases and misspecifications in available inflation measures are part of the problem, and there is a substantial literature on this subject. In contrast, theoretical specification of measurement concepts and empirical analysis of measurement problems for the variables that have been used on the righthand side of the price equation have received far less attention. For space limitations, I consider only one of the many variables that have appeared in price equations—wages.

A major theme of this paper is the disparity of existing research on the measurement of wages and prices, with the price measurement side having received much more rigorous and extensive analysis. This disparity exists even though conceptual specifications for wage measures and for consumer price measures are similar (a consequence of the fundamental analogy between the economic theories of production and consumption), and even though somewhat related bodies of studies exist on one or two of the major measurement problems. I argue that the analysis of inflation—and labor economics as well—would greatly benefit from extension to labor market variables of the type of research that has been carried out on the price measurement side.

I. Theoretical or Conceptual Specification[1]

I take as an axiom that economic measurements should be based on concepts drawn from economic theory. The principle can be applied at the micro level, or at the aggregate level.

In the case of consumer prices, the appropriate aggregate concept is the "constant-utility" price index, or "true cost-of-living index" (hereafter, *COL*), which measures the cost of remaining on the same indifference curve. The *COL* concept has undergone lengthy theoretical development, with important recent summary statements by Paul A. Samuelson and P. Swamy, and by Robert A. Pollak. The Consumer Price Index *(CPI)*, which is frequently used as a measure of inflation, is interpretable as a fixed-weight approximation to a *COL* (though differences other than weighting patterns may quantitatively be more important). Industrial price measures are equally important for inflation analysis and also require a conceptual framework; if they are used as *input* prices, their treatment is similar to that of wages, discussed below.

Deriving theoretical foundations for wage measurement concepts would appear at least as interesting as the *COL* problem, but has not captured economists' imaginations to the same degree. Wages serve labor economics as measures of income on the one hand and of the price of labor as a factor of production on the other. It is by now well established that the two uses imply different measurement concepts, but the income side need not concern us here, for in the analysis of inflation wages generally appear as employers' costs.

I presume that the question of interest takes the form: "How have wage changes affected production costs?" If labor is not a homogeneous input, so that different kinds must be combined (along with other inputs) in the production process, then this question implies a measurement concept closely analogous to a *COL*—an input price measure defined as the employer's cost (over time or space) of producing on the same production isoquant. We might

*Assistant Commissioner for Research, Methods and Standards, U S Bureau of Labor Statistics. Opinions are those of the author and do not necessarily reflect an official position of the Bureau of Labor Statistics. I am grateful to Robert A Pollak, Frank Stafford and Victor J Sheifer for helpful comments on an earlier draft

[1]Because of space limitations, and in response to a suggestion from the editors, I have suppressed specific references to the literature, retaining only those which are surveys or provide entree to the literature. Readers interested in specific points may consult references in the works cited.

expect, therefore, research on wage measurement concepts to proceed in a fashion parallel to existing research on the *COL*.

Two major theoretical issues in the price measurement literature are: 1) determining the appropriate form for the indexes, and 2) the question of "subindexes." Both questions turn on properties of the underlying behavioral functions, and both can also be asked of wage measures.

Duality theory shows that the form of the production or utility function dictates formulas for price and quantity indexes. For example, a Cobb-Douglas production (utility) function implies an input price index *(COL)* which is a weighted geometric mean of price relatives. Other production or utility functions yield other explicit price index formulas and some of them bear little resemblance to traditional price index forms. Most existing price measures are computed as fixed-weight indexes, frequently by Laspeyres' formula. Because it makes no allowance for substitution when relative prices change, a Laspeyres index provides an upward-biased measure of the base-period's *COL*—in fact, Laspeyres' formula is an upper bound on such a *COL*. A similar theorem shows that index forms which are inappropriate for the production function will produce biased measures of aggregate (economy wide or sectoral) wage change. Such specification bias could affect conclusions drawn about relations between wage and price movements.

On the price side, interest in the substitution bias in price indexes has led to empirical estimation of *COL*'s. *COL*'s can be computed using parameters from "complete" systems of demand equations, derived from explicit, multiproduct utility functions. A number of recent studies have estimated *COL*'s for the United States at a fairly disaggregated level, using a variety of utility function specifications. This research is summarized in Triplett. The various estimated *COL*'s all give very similar measures of inflation over the postwar period. Moreover, they all indicate that the specification error stemming from the fixed-weight property of conventional price indexes (that is, the fact that Laspeyres' formula cannot take into account substitution in consumption in response to changes in relative prices) amounts to only around 0.1 percent per year—a magnitude that is probably well within the appropriate confidence interval for the index.

Little analogous work on the production side has yet appeared. Though no economist, I take it, would consider a two-good utility function a satisfactory vehicle for serious research on consumer behavior, estimating production functions containing only the gross aggregates "capital" and "labor" has an honored tradition. Those few existing studies containing multiple labor categories have defined them at so high a level of aggregation ("white collar-blue collar," for example) that they have little relevance for measuring wages. Research on an input price analog to a *COL* requires a production function containing detailed occupations as labor inputs (plus other inputs at comparable levels of disaggregation). The required research is yet to be done, and until it is we will know very little about specification bias in input cost measures.

The subindex problem is equally interesting. A "subindex" is a component of the total index ("clothing," for example, or "meat"). The preceeding paragraphs have been written as if labor, materials, and capital costs were natural aggregates—which is in fact how they have invariably been treated in inflation studies and elsewhere. Such treatment amounts to the implicit assumption that each of these components can be formed as a subindex of the total production cost, or input cost, index.

Whether subaggregates of the full index are meaningful depends on properties of the underlying production or utility function, and not on the institutional considerations which have traditionally provided the rationale for the segregation of labor inputs from those of capital. In the production process, engineers may more appropriately be grouped with machinery than with janitors, so it is by no means certain that total labor inputs are an empirically appropriate

aggregate, or that employment costs are necessarily an appropriate subindex.

The theoretical requirements for consistent aggregation into subindexes have been extensively explored in both production function and consumer demand literature, but the empirical work that has accumulated is insufficient to draw many implications for the construction of either wage or price indexes.

II. Measuring the Micro Observations—Prices and Wages

At the micro level, issues which have dominated the price measurement literature have analogs on the wage measurement side. One is the question of quality error in price indexes.

A substantial body of empirical studies on the quality problem have accumulated and are surveyed in Triplett (pp. 30–61). Studies have produced "quality corrected" price indexes that differ by appreciable amounts from comparable indexes published by the Bureau of Labor Statistics *(BLS)* or other government statistical agencies There is, however, little agreement on the direction of the implied quality error.

> Some studies have produced "quality-adjusted" price indexes which rise more slowly than counterpart components in the *CPI* or *WPI* or deflators, at least for some time periods Other studies, however, suggest just the opposite—that in some components, and over some time periods, the investigator's indexes have risen more (or fallen less) than comparable *BLS* or *BEA* indexes. We are a long way from being able to estimate the quality error in the indexes; in fact, we are unable, in spite of the research that has been carried out on the problem, even to estimate the *sign* of the error in the overall indexes [Triplett, pp. 60–61]

Combining the results of existing studies to reach an estimate of the quality error in the aggregate indexes initially may seem an attractive approach. However, research indexes have usually been derived from a data base different from the one collected for the "official" index, and it has seldom been possible to decompose divergence in the two price indexes uniquely into quality measurement and data base, or sampling, components. Research on the quality problem is illustrative of its dimensions, but there are sufficient problems with most existing studies that one cannot be sure that substituting them for the official series would produce a superior price measure, especially since we lack an adequate concept for carrying out statistical tests of hypotheses about price index behavior.[2]

In sum, available research indicates that the quality problem in price indexes is a complex set of errors that are to a great extent unique to individual index components, and to particular pricing procedures, and which sometimes vary in sign from one time period to another. Moreover, the sign of the error in the index is *not* determined by the sign of quality change itself (improved quality, in other words, is consistent with either upward or downward quality error in the indexes). For more information, the reader is referred to my recent survey, and to the references cited therein.

[2]This is by far too complex a question to be explored here, but deserves mention because the problem has received so little attention Statistics on price index variance relate exclusively to the population from which the price quotations are sampled—that is, to one of the inputs to the process from which a price index ultimately emerges. I believe that the user is concerned with the total dispersion in measures of the *concept he wishes to employ*, so that a theory of index number variability should be constructed from the end-point of the process For example, if several equally defensible approximations to a true cost-of-living index give a variety of estimates, surely this is a component of dispersion that must be considered when making statements about the statistical significance of index change There are other procedural matters in the complex process of putting together a price index that are also subject to analyses as statistical processes.

The theory of sampling has been developed for cases where the uses are relatively simple (such as testing hypotheses about population means), in the price index literature, the sampling error notion has traditionally been supplemented with all kinds of *ad hoc* statements about so-called "procedural errors," as if the latter were not subject to analysis as statistical processes. What we need, and do not at present have, is a body of statistical theory for unifying all sources of price index dispersion into a comprehensive whole that would fully describe the dispersion in the index from all sources, from the point of view of the uses to which the final measure is put From this point of view, presently-published index "sampling error" is probably far too small, and is accordingly a misleading guide to the accuracy of the indexes

I know of no direct and explicit studies of quality error in wage measures comparable to those that have been carried out on the price side. Average hourly or weekly earnings are the most widely-employed wage measure for the analysis of inflation. These data are analogous to unit value indexes on the price measurement side, as they are formed by dividing establishment payrolls by measures of hourly or weekly employment. It is easy to compile a list of references acknowledging the shortcomings of average earnings statistics as wage data. Exclusion of fringe benefit payments, and sensitivity to employment mix changes and overtime schedules are measurement errors which are frequently cited in the labor economics literature and by the agencies which compile and release the data.

Recently (June, 1976) the *BLS* inaugurated a new wage index, designated the "Employment Cost Index," or *ECI*. At the micro level the *ECI* measures employers' hourly labor compensation costs for specified occupational categories and combines observations into an index having fixed employment weights, using methods comparable to those of the well-established *CPI*. Data from the *ECI* should prove a substantial conceptual advance over alternative wage measures which have been employed in the analysis of inflation, but the labor quality problem will still demand attention

The wage differential literature provides perhaps the most careful examination of wage measurement questions. A large part (up to two-thirds) of the apparent difference in wage levels between regions is accounted for by various characteristics of the employees or of the employing establishment or its location (age, sex, race and education of employees, city and establishment size and industry have all been identified as important elements). Comparable results emerge from analysis of interindustry and other differentials. Though the variables cited are all interpretable as bias when the data are used for the analysis of inflation, one expects such sources to create smaller discrepancies in intertemporal earnings comparisons, and they should be smaller still in components of the *ECI*, which was designed to eliminate many of the problems with average hourly earnings data.

Clearly, the human capital literature is the starting point for studies of labor quality. One portion of the human capital literature has focused on rate of return (to educational investment) calculations, which are relevant for analyzing labor supply questions (such as occupational choice and training decisions), but not for the present discussion. Another orientation employs labor quality measures to improve some economic measurement (for purposes such as analyzing the sources of economic growth). Research techniques in this part of the human capital literature are directly analogous to "hedonic" methods, which have been the primary vehicle for research on quality measurement on the product side

Making use of the analogy, we may define a hedonic price (wage) function as a relation between a cross-section of prices (wages) and a set of product characteristics (worker characteristics); the characteristics selected as independent variables are usually presumed to be the true arguments of the utility (production) function. In the consumption case, this means that the parameters of the hedonic price function are interpretable as establishing consumption opportunity loci on characteristics, in the manner popularized by Kelvin Lancaster.

Despite the wide professional acceptance of the human capital concept, I remain pessimistic about its potential for making labor quality adjustments in wage measures. When hedonic wage equations are run *across* the occupational structure, a reasonable amount of wage variance is accounted for by human capital variables. But adjustments from interoccupational hedonic wage equations are useful mainly when the wage or employment measures being adjusted are gross aggregates (total hours or employment, industry-level average earnings, and the like), which are greatly affected by employment shift effects. The employment shift problem can more readily be attacked through im-

proved data collection procedures; the problem should be greatly ameliorated in the new *ECI* data, because the unit of observation is the occupation, and occupational weights are held constant in deriving the industry wage measures.

Thus, what we need are adjustments for labor quality *within* occupations, and here hedonic wage equations show far less explanatory power. Wage equations on micro data seldom produce R^2 values in excess of .3. In contrast, hedonic price studies exhibiting R^2 values of .9 and above are common. Unless the unexplained variance in hedonic wage equations is unrelated to labor quality, adjustments based on within-occupation regressions are not likely to account for very much of the quality problem in wage measures, and for this reason hedonic wage equations seem a less promising tool for measurement problems than hedonic studies have proved to be on the price measurement side.

Theoretical research on the conceptual framework underlying hedonic measures and on the economic justification for their use has proceeded mostly on the price measurement side (even though the human capital literature is far larger than the literature on hedonic studies). In contrast, technical discussion in the human capital literature concerns mainly estimation methods, not the basic economics, and most of the real controversy over the use and interpretation of the results amounts to an attack on the assumptions of neoclassical economics (for example, denial that the education-earnings relation is productivity related).

All is not well on the hedonic side, however, as some of the hedonic framework literature has confused the problem of interpreting hedonic measures with questions concerning their appropriate use. For a hedonic function to have economic meaning, one must assume that the characteristics are the arguments of the utility function, rather than those aggregations of characteristics we conventionally call "goods." By extension, a hedonic wage equation rests on the assumption that its independent variables (worker characteristics) are the productive factors that serve as inputs in the production function. Note that this is a stronger condition than the assertion that the characteristics "explain," or "determine," some unobserved scalar measure called "quality." The assumption does not imply, however, that the hedonic function is to be *derived* from the utility function, because the hedonic function provides the characteristics space analog to the familiar budget constraint. It is thus an estimate of the price structure the consumer faces, and not the objective function he is maximizing. This view of the hedonic function does imply that hedonic estimates are valid for use as adjustment factors in price indexes only when subindexes defined on the characteristics of the hedonic function are legitimate, which is a serious theoretical limitation. Comparable limitations apply to the human capital, or hedonic wage, literature, but have not been developed there.

As a concluding comment on micro measurement, I note that the most detailed wage data published regularly are arrayed by *industry*. For estimating price equations, or analyzing production functions, industry groupings are indeed appropriate. A great deal of the empirical literature on inflation, however, has concerned itself with the analysis of wage inflation (see the survey by David Laidler and Michael Parkin). For the analysis of labor markets, data grouped by industries frequently are not relevant, unless the investigator is content to work with very high levels of aggregation, or to restrict attention to those relatively small number of cases where industry and labor market are coterminous. Though much empirical work in labor economics has used data on industry wages to explore propositions about labor markets, I suspect that this lack of concern about the inappropriateness of data has something to do with the conclusion (which seems endemic among some labor economists) that the theory of markets contributes little to explaining labor market phenomena. It is discouraging that so little attention has been paid to measurement questions in the literature on wage determination.

III. Conclusions—Constructing a Framework for Measurement

One cannot claim that the price measurement literature has solved all of its own problems. I have used it as a model only to demonstrate how far behind is scientific work on the question of wage measurement. Much of the subject of labor economics has been devoted to explaining wage patterns, yet it is hard to find in this literature any clear and rigorous definition of the measurement concept to be explained. A similar deficiency exists in the inflation literature. the Laidler and Parkin survey does not even mention measurement problems, an omission which merely reflects what the empirical literature on inflation has contained on this subject. Effective analysis requires clear specification of the nature of the data that are appropriate to the hypotheses being considered, and careful consideration of the effects of deviations between what one wants by way of data and the data which are actually available.

In view of the findings of labor market research over the past decade or so, constructing a conceptual framework for measuring employment costs will not be a simple task. Much of our wage data implicitly rests on a framework derived from the older collective bargaining tradition of labor market studies. Fifteen years ago, the theory of labor costs would still have been a theory of wage *rates* (perhaps modified by inclusion of "fringe benefits"); recognition that the employment contract involved much more than the purchase of an hour's labor was relegated to the institutional detail. In the interim, theoretical and empirical work has brought an awesome array of factors into the theory of the firm's behavior toward its labor input—the search literature, recognition of the "quasi-fixed" factor status of some labor inputs, the labor market dynamics literature, and the analysis of "effective" labor input are only some of these factors.

Many strands of recent research on labor markets may be characterized as attempts to make more realistic specifications of the firm's employment costs. We derive our conception of what we want to measure in the wage area from precisely this subject—the specification of the nature of employment costs, and the theory of the firm's behavior toward its labor input. A modern conceptual framework for wage measurement can therefore only proceed from a thorough examination of the implications of the entire body of recent labor market research.

Unfortunately, the task of distilling the wage measurement implications out of the revolution in labor economics that has occurred over the last fifteen years has hardly—if at all—begun. Though it is clear that elaboration of labor market theory has resulted in a more powerful theory, one that enables us to explain a far larger proportion of labor market behavior, yet it greatly complicates the problem of determining how employment costs are to be measured. The solution will not be easy. Moreover, it is not a task that one can expect to be carried out by the data producers alone. Development of an "Economics of Measurement" proceeds hand in glove with economic research Too little attention has been paid to measurement concepts in labor market research.

REFERENCES

David Laidler and Michael Parkin, "Inflation: A Survey," *Econ. J.*, Dec. 1975, *85*, 741–809.

Kelvin Lancaster, *Consumer Demand: A New Approach*, Columbia University Press 1971.

Robert A. Pollak, "The Theory of the Cost of Living Index," Bureau of Labor Statistics Working Pap., no. 11, 1971.

Paul A. Samuelson and S. Swamy, "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis," *Amer. Econ. Rev.*, Sept. 1974, *64*, 566–93.

Jack E. Triplett, "The Measurement of Inflation: A Survey of Research on the Accuracy of Price Indexes," in Paul H. Earl, ed., *Analysis of Inflation*, Lexington Books 1975.

An Integrated Model of Final and Intermediate Demand by Stage of Process: A Progress Report

By JOEL POPKIN*

The purpose of this paper is to report on progress in developing an integrated model of final demand and intermediate output. The model was conceived to provide a framework for studying the transmission of inflation through the U.S. economy. Because price movements are interrelated with the behavior of other economic variables, such as output and inventories, the inclusion of these variables in the model permits analysis of transmission effects for them as well.

I. A Description of the Model

The origin of the model is a stage-of-process price model of the U.S economy linking movements of the crude, intermediate and finished goods components of the Wholesale Price Index (*WPI*) with each other and with the Consumer Price Index (*CPI*). By comparison with the more widespread analysis of prices by standard industrial classification, a stage-of-process framework is demand oriented.[1] Aggregation is determined by the nature of the purchaser, highlighting the direction of shipments and providing information on the determinants of changes in demand and the timing of their impact However, aggregation by stage-of-process presupposes that the input-output structure of an economy is triangular. If not, then the flows are probably so complex that they can be analyzed only by considerable disaggregation, probably to a level at which there would be a paucity of the monthly or quarterly data on prices and their determinants needed to capture the dynamics of inflation and its transmission. It turns out that the 1967 input-output table for the United States at the 486 order level can be aggregated into a 37 order table in which only about 5 percent of the transactions (column totals) lie below the main diagonal Time series data have been developed for the commodity-producing cells among the 37 These data are used to construct a model of the intermediate economy which is linked to a model of final demand.

Figure 1 depicts the basic structure of the model. This figure has been simplified for illustrative purposes to show primarily the flows among three sectors—a final commodity demand sector, the finished goods manufacturing sector that supplies it, and one of the primary manufacturing industries that produces materials required to manufacture the finished goods.[2]

There are several blocks of equations in the model of the intermediate economy:

1) New orders received. These depend on a proxy for expected sales and the difference between actual and desired inventories for the sectors placing the orders, and the ratio of the price of goods being ordered to the price at which the commodities made from them are being sold. The assumptions regarding the source of orders are based on the 1967 input-output table and are simplified because that table is nearly triangular.

2) Unfilled orders, production, and inventories of finished goods and goods-in-process.

*National Bureau of Economic Research The research reported in this paper was supported by a grant from the National Science Foundation Michael McKee, David Crary and Joslin DePuy provided valuable assistance

[1]Stage-of-process classification is appropriate and useful also in the study of supply side relationships, in particular of the production function An important economic issue is whether capital and natural resources are substitutes or complements Classification of industries by the intensity of their raw material usage would yield sector groupings quite similar to those that result from stage-of-process analysis, sectors for which three factor (capital, labor, materials) production functions could be estimated

[2]In Figure 1 the *K* signifies intersectoral influences within the trade block, with *L* indicating a lagged effect

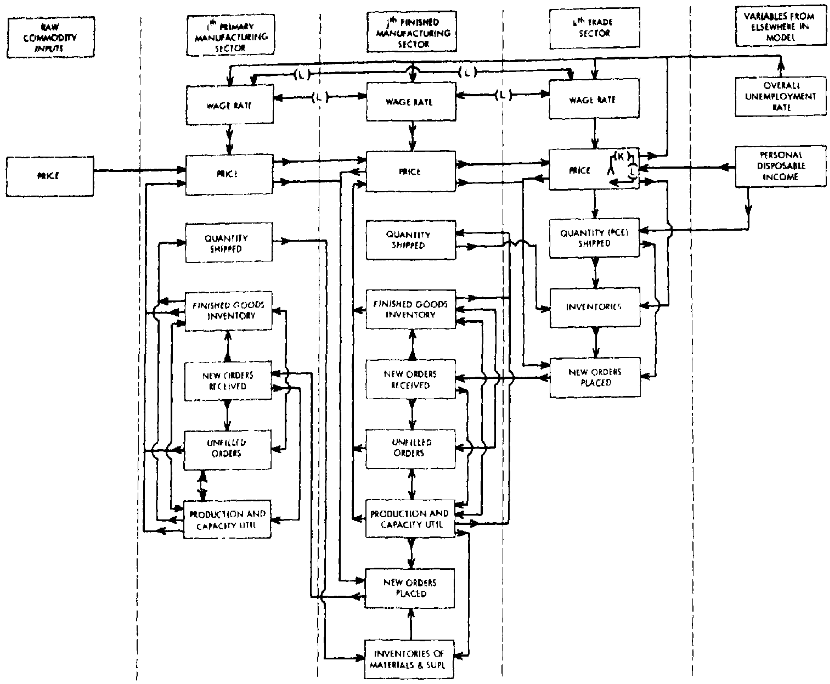


FIGURE 1. FLOW DIAGRAM—INTEGRATED MODEL OF FINAL AND INTERMEDIATE DEMAND BY STAGE OF PROCESS

These equations are found in the recent work of Gerald Childs and are based on the application of linear decision rules at the firm level in the simultaneous determination of inventories, production, and unfilled orders. Equations for each of these three variables, estimated for the stage-of-process sectors in manufacturing, contain current new orders as a proxy for expected orders and the lagged values of the three dependent variables. Only two of the three equations need to be estimated; the third variable can be solved through an identity.

3) Prices (of orders or sales) The structure of the price equations reflect the results of testing various hypotheses about their determination. Variable input prices—materials and labor—are common to the price equations for every sector. Various combinations of other variables—new orders, unfilled orders, inventories (finished goods and goods-in-process), production and productive capacity—enter significantly in most of the other equations. The significance, or insignificance of these variables permits inferences to be drawn about the structure of markets. Those variables that are significant are generated by the equations in blocks 1 and 2 just described. No distinction is made at present between producers and purchasers prices; producers prices are used throughout.

4) Wage rates. Wage rate series for each sector have been constructed from average hourly earnings data adjusted for overtime in manufacturing and interindustry shifts. Such data have been further adjusted to reflect fringe benefits based on annual surveys of labor costs. These variables for each sector are explained by

some combination of the national unemployment rate, a measure of economic activity in each specific industry, the deflator for personal consumption expenditures (*PCE*) and a variable reflecting past behavior of each industry's wage rate relative to those in other industries.

In addition to these four blocks of equations there are several sets of identities, exact or estimated, which provide values for such variables as shipments and materials and supplies inventories in each industry.

The final demand model to which the intermediate model is being linked resembles, except for some differences to be mentioned shortly, the typical Keynesian macro model in use today; it is however substantially smaller—50 or so equations and identities. The size of the full model—intermediate and final demand—is however about the same as many final demand models in use today An obvious assumption implicit in the work reported here is that at least for the analysis of inflation, it is better to use limited resources to model both intermediate and final demand sectors rather than to disaggregate final demand itself. This assumption appears to be merited based on the predictive ability of a set of stage-of-process price equations (Popkin 1974). Whether it holds in the context of a full model remains to be evaluated.

There are three respects in which the final demand model employed in this study differs from those currently in use. First, there is no aggregative inventory change equation. Such change is based on an estimated identity linking inventory changes occurring in each specific intermediate and final demand sector of the model. Second, the categories into which *PCE* is disaggregated differ somewhat from those published quarterly by the Bureau of Economic Analysis (*BEA*) in order to achieve a better concordance between final demand sectors and the intermediate industries that supply them. A third distinction is that the various final demand deflators are estimated for each sector as a function of variables, including materials prices relating to that sector; these deflators are then added to obtain the *GNP* deflator. In some current models the *GNP* deflator is derived first and then used along with other variables, to explain the component deflators of final demand.

II. Preliminary Results for Some Specific Variables

A. The Price Block

Price equations have been estimated for twenty-one sectors, sixteen that manufacture commodities and five that distribute them.[3] There are eight primary manufacturing sectors, defined as the first processors of raw materials other than food or fuel. The eight are textiles, lumber, paper, chemicals, fertilizers, steel, nonferrous metals, and stone, clay and glass. There is one semifinished manufacturing sector which is not in the basic model but for which a price equation was estimated nonetheless.[4] Price equations are estimated for seven finished goods industries, five of which produce primarily consumer finished goods. The five are consumer staples (excluding food and fuel), consumer home goods, automotive, petroleum and consumer food, beverages and tobacco. The other two are producer goods industries, machinery and nontransportation equipment, and a grouping of industries producing ordnance, ships, aircraft and railroad equipment, much of which is purchased by the federal government. For the five sectors producing con-

[3] The results are based on monthly Census data on inventories, shipments, and new and unfilled orders that have been benchmarked only through 1971, and Federal Reserve Board production indexes and National Accounts data of the *BEA* prior to their revision in 1976 The author wishes to thank these agencies and the Bureau of Labor Statistics for their cooperation in providing data for the project

[4] Semifinished manufactures, those that are neither primary nor finished commodities, are quite heterogeneous. The purchases by finished goods industries of semifinished commodities were allocated to the primary industries on the basis of their shipments to semi-finished goods industries Thus all commodities with steel content purchased by the auto industry from the steel industry directly or from a semifinished good industry are assumed, based on their steel content, to have been purchased from the steel industry. The semifinished goods industry is assumed to produce only value added in the full model

sumer finished goods, retail price equations are estimated, linking manufacturers prices of such goods to *PCE* component deflators, made up of *CPI* prices of similar goods.

Several price equations have been estimated for each of the sixteen, stage-of-process manufacturing sectors for two sample periods in order to test various hypotheses about price determination.[5] These hypotheses, the specifications to which they give rise and the results obtained for the eight primary producing industries are reported elsewhere (Popkin 1976). Space precludes more than a summary of the results here. Without going into the implications of these findings for market structure they may be divided into two categories: 1) those that suggest that the ratio of output prices to the weighted sum of labor and materials' prices behaves procyclicly, and 2) those in which the ratio behaves anticyclicly or is cyclicly neutral. In seven of the sixteen manufacturing sectors the ratio behaves procyclicly in both of the sample periods for which the analysis was conducted. Four of those seven cases are primary producing sectors. They are textiles, lumber, paper, and nonferrous metals. The other three are the semifinished manufacturing sector, petroleum manufacturing and machinery and equipment.

The ratio does not behave procyclicly in either sample period in five manufacturing sectors. Three of these are the rather sizeable industries producing consumer staples (other than food and fuel), consumer home goods, and autos.[6] The other two are the steel industry, and the ordnance, ship-building, aircraft and railroad equipment sectors.

Pending the development of inventory data based on the recent revisions of the National Accounts, it is not possible to thoroughly test hypotheses concerning the output-input price ratios for the five sectors distributing consumer finished goods. Preliminary tests based on excess-demand proxies like the unemployment rate suggest the ratio behaves procyclicly in the sectors distributing automobiles and consumer staples (other than food and fuel). The ratio does not appear to behave procyclicly in the sectors distributing food, home goods and petroleum products.

When the results for the twenty-one sectors are viewed as a whole, they appear to support the general conclusion that it is in the primary and semifinished goods industries, rather than in the finished-goods-producing and distributing industries in which demand influences the relationship of output to input prices.[7] This is particularly apparent for consumer goods.

Perhaps related to this finding is one by Willard Mueller and Larry Hamm that while the four-firm shipments concentration ratio for total manufacturing has drifted up only .8 percentage points between 1958 and 1970, the ratio for consumer goods industries shows an increase of 2.6 percentage points. Eight-firm concentration ratios calculated for the sixteen stage-of-process manufacturing industries in the model reported in this paper show similar behavior. From 1958 to 1972, latest data available, concentration in consumer finished goods industries increased by 5.1 percentage points, compared with a rise of 1 2 percentage points for producer finished goods industries and of .2 percentage points for producers of primary and semifinished manufactures. Further work is required to test whether such changes in concentration ratios have a causal role in the finding reported above that rates of change of prices of consumer finished goods are less sensitive to changes in demand than are price changes for primary and semifinished manufactures.

Regardless of the reason for the finding, it suggests the source of what is regarded as insensitivity or considerable delay in the response of prices to a change in aggregate demand policies, particularly a restrictive change. Accord-

[5]The sample periods are 1959–71 (2nd qtr) and 1959–75 The first sample period excludes observations subsequent to the initiation in August 1971 of the wage and price control period

[6]For autos data are available for the shorter sample period only.

[7]Charles L Schultze noted similar behavior in his analysis of the nature of inflation in 1955–57

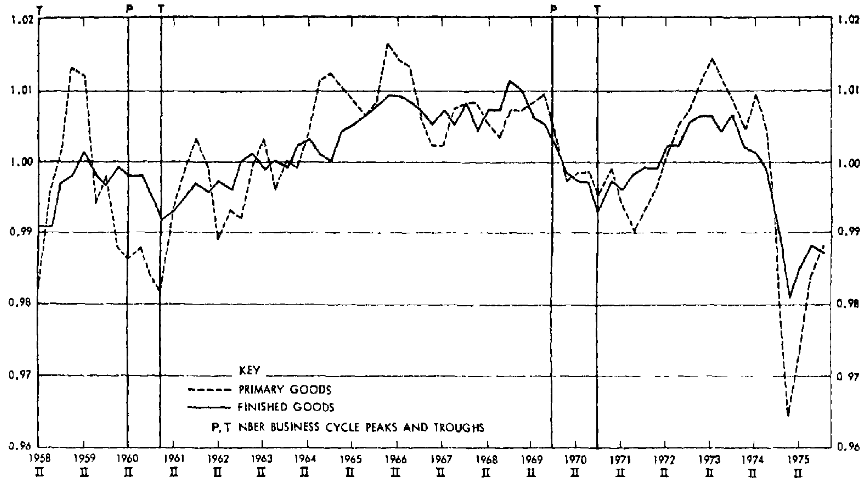


FIGURE 2 PERCENTAGE DEVIATIONS FROM TREND IN NEW ORDERS RECEIVED BY PRIMARY AND FINISHED GOODS MANUFACTURING INDUSTRIES

ing to this analysis the initial response to such a change on the part of most retailers and manufacturers of consumer finished goods is to reduce output, not the ratio of output to input prices. The reduction in the volume of retail sales and finished goods production is greater than would be the case if the output-input price ratio behaved procyclicly. This results in larger cutbacks in orders for materials and supplies placed by finished goods producers than would occur otherwise. When these cutbacks in orders impact on those semifinished and primary manufacturing industries in which the output-input price ratio does behave procyclicly, prices in these industries weaken. Such weakness then feeds forward to final demand prices, but, of course, with a lag, affecting prices in all finished manufacturing and distribution sectors, whether or not the output-input price ratio behaves procyclicly.

B. The New Orders Block

Figure 2 contains two new orders series (in 1967 dollars) plotted quarterly from 1958 through 1975; both series are expressed as percentage deviations from trend. One series is orders received by finished goods manufacturers, the second, orders received by primary manufacturers. The second series fluctuates with greater amplitude than the first; it rises more rapidly during expansions, particularly the early phase, and falls faster during contractions. The tendency for output (and prices) to fluctuate more at the earlier stages of production than at the later stages (through distribution to final purchases) has been noted in the research of others, particularly that undertaken at the National Bureau of Economic Research.[8] It is reasonable to expect the same pattern applies to new orders.

Such behavior can be understood in terms of a relationship in which new orders at each stage of production and distribution depend on expected sales and the difference between actual and desired inventories.[9] The difference between actual and desired inventories has two

[8]Particularly Frederick Mills, Moses Abramovitz, Ruth Mack, Geoffrey Moore, and Victor Zarnowitz.

[9]Prices of both the orders and the products that will be produced from the materials being ordered will be ignored for the moment, but belong in the orders function.

components: one arises as a result of any error made in anticipating current sales while the second reflects the marginal adjustment to inventories required as a result of a change in expectations about future sales, its size being based on the inventory accelerator—the desired ratio of inventories to sales. For simplicity assume that orders are placed at the end of a time period and satisfied immediately from stocks and that last period's sales (or orders received) are proxies for expected sales in the current period Then:

$$(1) \qquad O_{f,t} = X_{f,t-1} + (X_{f,t} - X_{f,t-1}) + \alpha(X_{f,t} - X_{f,t-1}),$$

where O_f is orders for finished manufactures placed by retailers, X_f, retail sales and, α, the desired inventory-sales ratio of retailers. A similar relationship can be developed for orders for primary materials (O_p) placed by finished goods manufacturers.

$$(2) \qquad O_{p,t} = O_{f,t-1} + (O_{f,t} - O_{f,t-1}) + \beta(O_{f,t} - O_{f,t-1}),$$

where β is the desired inventory-sales ratio of finished goods manufacturers Simplifying (1) and (2) yields:

$$(3) \qquad O_{f,t} = X_{f,t} + \alpha(\Delta X_{f,t}) \text{ and}$$
$$(4) \qquad O_{p,t} = O_{f,t} + \beta(\Delta O_{f,t}).$$

From (3) orders placed by retailers for finished manufactures fluctuate with greater amplitude than retail sales (as a result of the term $\alpha\Delta X_{f,t}$) and from (4) orders received by primary manufacturers fluctuate with greater amplitude than those received by finished goods manufacturers (as a result of the term $\beta\Delta O_{f,t}$).

When (3) is substituted into (4) the result is

$$(5) \qquad O_{p,t} = X_{f,t} + (\alpha + \beta)(\Delta X_{f,t}) + (\alpha\beta)(\Delta^2 X_{f,t}).$$

An understanding of the differences in the trend deviations of orders of finished and primary manufacturers depicted in Figure 2 is gained by subtracting $O_{f,t}$ from $O_{p,t}$ which yields:

$$(6) \quad O_{p,t} - O_{f,t} = \beta(\Delta X_{f,t}) + \alpha\beta(\Delta^2 X_{f,t}).$$

This model suggests that when economic activity is increasing at an increasing rate (as it usually does during the early phases of an expansion) or is declining at an increasing rate (the early phase of a recession), O_p will rise or fall, faster than O_f.[10] This result appears to be consistent with the data in Figure 2. However, more testing is required before the behavioral hypothesis described above can be accepted.

III. Concluding Remarks

This paper has been intended to demonstrate that it is feasible to develop the data required to construct an integrated model of intermediate and final demand. Based on work done so far for prices it appears that such a model will improve understanding and predictability of inflation, at least in a partial equilibrium sense. There is much to suggest also that the full model will be useful Time limits the coverage that has been accorded to results for other blocks of equations; so far, on a single equation basis they look promising. However judgment on the issue must await full model simulations and forecasts.

The conceptual and empirical work presented here does appear to have implications for economic stabilization. It has been shown, under seemingly defensible assumptions, that the amplitude of cycles in output (orders) increases as one looks behind those of final demand to those of finished manufactures and then to primary

[10] Using the assumptions posited here, Michael McKee has developed the proof for this statement in the case where X_f fluctuates regularly in a sinusoidal pattern with a cycle length approximating the length of the relevant business cycle This proof establishes the conditions under which O_p, or the ratio O_p/O_f, may be a leading cyclical indicator

production. Empirically it has been established that the ratio of price to unit variable costs is relatively unresponsive to demand at the stages of manufacture and distribution of finished goods, particularly nonfood, consumer goods, and that the responsiveness of that ratio is larger in primary manufacturing industries. Taken together this means that changes in final demand result in larger changes in output than would obtain if the ratio were more responsive at later stages of production and distribution; and that it takes longer than would otherwise be the case for changes in final demand to affect that part of inflation represented by the difference between prices and unit variable costs.

REFERENCES

Moses Abramovitz, *Inventories and Business Cycles, with Special Reference to Manufacturers' Inventories,* New York 1965.

Gerald L. Childs, *Unfilled Orders and Inventories: A Structural Analysis,* Amsterdam 1967.

Ruth P. Mack, *Information, Expectations, and Inventory Fluctuation: A Study of Materials Stock on Hand and on Order,* New York 1967.

Frederick C. Mills, *Price-Quantity Interactions in Business Cycles,* New York 1946.

Geoffrey H. Moore, "The Cyclical Behavior of Prices," in Victor Zarnowitz, ed., *The Business Cycle Today,* New York 1972.

Willard F. Mueller and Larry G. Hamm, "Trends in Industrial Market Concentration, 1947 to 1970," *Rev. Econ. Statist.,* Nov. 1974, 56, 511–20.

Joel Popkin, "Consumer and Wholesale Prices in a Model of Price Behavior by Stage of Processing," *Rev. Econ. Statist.,* Nov. 1974, 56, 486–501.

———, "Price Behavior in Primary Manufacturing Industries, 1958–73," National Bureau of Economic Research Working Paper No. 136 (unpublished), 1976.

Charles L. Schultze, "Recent Inflation in the United States," Study Paper No. 1 for Joint Economic Committee in connection with their Study of Employment, Growth, and Price Levels, Washington, D.C. 1959.

Victor Zarnowitz, *Orders, Production, and Investment: A Cyclical and Structural Analysis,* New York 1973.

The Explanation of Inflation: Some International Evidence

By KARL BRUNNER AND ALLAN H. MELTZER*

Explanations of inflation can be subdivided into two major groups. "Sociological theories" assert that movements of prices and wages proceed independently of market conditions. Economic theories on the other hand elaborate the essential dependence of price-wage movements on evolving market conditions. Sociological theories, dominant in Europe, emphasize the role of a wide array of institutional arrangements. In contrast, according to economic theory all forces and events affect inflation via market processes.

Within the class of economic theories, there are substantial differences in emphasis. Many of the differences are about the nature of dominant impulses. Which forces or actions produce disturbances that are systematically related to inflation? Many policy disputes have as their central, intellectual issue, the nature of the impulses generating inflation and unemployment.

Most recent discussion in economics has been about the details or properties of the mechanisms by which inflation is transmitted from one market to another. Such studies are valuable, but they cannot resolve policy disputes about the relative importance of financial and real disturbances in the generation of inflation. This paper is addressed to the material involved in such disputes. It reports on work in progress by a group concerned with the comparison of the relative importance of various impulses in the inflationary process. Observations drawn from five countries are used to assess the major issues bearing on crucial questions of policy.

I. General Remarks on the Class of Economic Theories

To distinguish the inflation problem from other adjustments of the aggregate price level, it seems useful to partition the observed relative change of the price level $\hat{p}$ into two components π and ρ, where $\hat{p} = \pi + \rho$. The first component, π, refers to a persistent and sustained increase in the price level, and ρ summarizes the many passing effects involving once and for all adjustments in the price level. Price theory informs us that the price level reflects an interaction between financial and real conditions. Changes in the price level result from changes in underlying real or financial circumstances modifying the general market conditions of an economy. An application of price theory therefore, directs attention to both real and financial factors. Further differentiation in the analysis emerges at this stage from different evaluations of the relative role of real and monetary factors in the observed adjustment of the price level. These alternative evaluations yield alternative approaches *within* the price-theoretical approach.

One thesis advances an eclectic view. The price level adjusts to a continuous series of erratic events or random changes in real or financial conditions. Any change or event is equally possible or probable and their combined evolution moves the aggregate price level over time. This thesis assigns no significance to π beyond its representation as a statistical average. The movements of $\hat{p}$ coincides essentially with ρ, and the inflation problem is interpreted as a

*University of Rochester and Carnegie-Mellon University, respectively. This paper forms part of projects supported by grants from the National Science Foundation to both authors. We gratefully acknowledge the contribution of our co-workers in the International Monetarist Consortium Dean Dutton, Michele Fratianni, André Fourcans, Pieter Korteweg, Johan Myhrman, and Manfred Neumann. They should not be entirely absolved from blame or credit for this paper.

sequential short-run adjustment of the price level from period to period.

An alternative thesis emphasizes the occurrence of systematic and dominant impulses. This thesis seems to us a more promising avenue for the development of a useful explanation of inflation. Two steps are required for the explanation. The first interprets the nature of dominant impulse patterns, and the second relates these patterns to the decomposition of the observed rates of price change, $\hat{p}$.

Financial shocks emitted by fiscal actions and monetary events are among the dominant impulses. Occasionally, the literature refers also to autonomous movements of the anticipated net yield on real capital, the Keynes-Wicksell impulse. We believe that this impulse can be discarded from the list of systematic forces The consequences of its operation are inconsistent with the observations made on asset prices and investment spending in Italy, Germany, United Kingdom and other countries where real returns declined with inflation We also discard the operation of systematic real shocks from the list of dominant impulses. Experiences of the past three years show that real shocks occur, but they do not occur in the manner required to explain a *persistent* increase in the price level. Real wealth and real income would have to fall continuously to satisfy this requirement.

The occurrence of real shocks, including Wicksellian impulses, is related to the ρ component. Real shocks induce once for all adjustments in the price level and explain transitory changes in both directions. The persistent and sustained movements of the price level, the π component, are attributed to the financial impulses. The issue can be described in a diagram in the price-output plane, juxtaposing an aggregate demand curve with an aggregate supply curve.[1] Expressed in diagrammatic terms, the hypothesis states that the position of the supply curve adjusts, beyond the short run, to the movements of the demand curve. Adjustments of expectations and the interaction between wages and prices, or labor and output market, link the position of the supply curve over a longer run to variations in the position of demand.[2] We may note in passing that many aspects emphasized in sociological explanations can be naturally reinterpreted and integrated into the account outlined above.

Our account has been restricted so far to a closed economy and is thus applicable to the world as a whole. An examination of the inflation problem confronting single countries enmeshed in international transactions introduces additional issues, notably the relative role of domestic and external factors in domestic inflation.

Our analysis specifies four channels conveying external influences. Two channels operate on aggregate demand, and two others simultaneously on supply and demand. Inflation abroad produces a balance of payment surplus and raises export demand. The latter affects aggregate demand for domestic output immediately. The payments surplus raises the monetary base and also affects aggregate demand by changing interest rates and asset prices. But foreign inflation also raises import prices, including the prices of inputs to domestic production. Domestic wages also respond to rising import prices. The rise in input prices and wages shifts the supply curve. But changes in input prices induce substitutions of domestic for foreign products and increases in wages increase aggregate demand. An open economy with a fixed exchange rate system experiences in this manner some relatively autonomous influences operating independently of its own conditions. The role of these autonomous external impulses has been much emphasized in European literature (particularly by John R. Hicks) and in American discussions in past years.

The influence of external effects on domestic inflation deserves investigation. Our previous

[1]These curves were described in several articles The reader should note that the aggregate demand curve is really a semireduced form. A point on the curve represents (p-y) combinations satisfying the output *and* the asset market equations Properties of asset markets are thus impounded into the demand curve

[2]The length of this run depends on the public's assessment of future policy patterns based on its past experiences.

argument implies, however, that we cannot explain world inflation by reliance on simultaneous operation of "external effects." The external effects relative to a given country reflect the already existing inflation outside the country, and this inflation is reducible, according to our hypotheses, to the trend in financial impulses the world over. A study of individual countries with different economic weight and "location" in the world economy may still offer some useful information on the state of our basic hypothesis and several related aspects and can help to establish the relative weight of domestic financial impulses and external effects.

II. A Summary of the Preliminary Findings

A comparison of the relative importance of various impulses in the generation of inflation in the United States, Germany, France, Italy, The Netherlands, and Sweden is a principal aim of our study. One of the first tasks was the measurement of the financial impulses and of the external impulses. A fiscal impulse was computed for each country. The fiscal impulse, *FI*, is a weighted combination of relative changes of government expenditures and tax revenues, with proper adjustment of both expenditures and revenues for feedback effects from prices and real income. The detail varies somewhat from country to country, reflecting the particular country's circumstances. The relative change of the money stock (usually M_1) was used as a measure of the monetary impulse, *MI*. Two comments are needed in this context. First, the monetary impulse contains both domestic effects and also one channel (via the balance of payments) of the external effect. These separate effects embedded in monetary growth can be examined by "going behind the money stock." Second, monetary growth was not adjusted for feedbacks through the balance of payments accounts. These feedbacks introduce a negative relation between domestic inflation and monetary growth over the sample period (with fixed exchange rates) considered here. Unadjusted monetary growth, containing these feedbacks, cannot bias the statistical results in favor of the monetary impulse.

Two external effects were measured, *EI(P)* summarizes the price impulse via imports, and *EI(Q)* expresses the direct impulse via expanding foreign demand for domestic output. Once again, some unavoidable differences occurred in the detailed construction of the two external impulses for the various countries examined.[3]

The method of evaluation uses a nonparametric and a variety of parametric methods. Tables 1 and 2 summarize a preliminary set of nonparametric tests bearing on alternative, narrowly conceived dominant impulse hypotheses. Each one of the four impulses is subjected to an examination as a dominant impulse with respect to changes in the price level (usually represented by a consumer price index) and with respect to a measure of output. Even if one is inclined to admit, as our analysis implies, that all four impulses are expected to exert simultaneous influence on changes in the price level and output, the comparison of the single impulse tests yields some useful information about broad orders of magnitudes operating over the sample periods. The test is based on a three-by-three contingency table with rows and columns distinguished according to the sign of the variables involved (+, −, 0). First differences of the impulse measures were required for the test, i.e., acceleration of the price level (first difference $\Delta\hat{p}$ of $\hat{p}$) and similar first differences of *FI*, *MI*, *EI(P)* and *EI(Q)*. All cases are based on annual data. Some results (e.g., The Netherlands) use concurrent values of the variables, whereas in others (e.g., Germany) the results are based on lag patterns yielding the best results for the variable under consideration. The reader should be warned not to interpret a negative ∇p statistic as reflection of a negative association. The negative sign of the nonparametric statistic simply reflects dominant effects via the off-diagonal cells in the contingency table. The negative sign reveals a dominant negative *or* zero association.

[3] The detailed measures and procedures are described in the papers prepared by the authors for the respective countries. These papers will be publicly available at the Carnegie-Rochester Conference on Public Policy, April 1977, at the University of Rochester.

TABLE 1—THE ∇p STATISTIC MEASURING THE ASSOCIATION BETWEEN CHANGING IMPULSES AND ACCELERATIONS OF THE PRICE LEVEL

	United States	Netherlands	Germany	Italy	France
FI	-2.6	6	6.7	17	.04
MI	2.0	3.3	23.3	46*	.35*
EI(P)	-4.8	2.3	52.1	56*	09
EI(Q)	8	8	19.4	38*	-01

[a]The ∇p statistic has been standardized for the United States, Germany and The Netherlands. It is not standardized for Italy and France

[b]The significance levels for the standardized statistic are 1 65 for 5 percent, and 2 33 for 1 percent.

[c]The stars in the columns describing Italian and French statistics indicate statistical significance at the 1-%-level

[d]The sample periods are United States, 1952–72 (annual data) The Netherlands 1953–73 (annual data) Germany quarterly data IV/1953–I/1974 Italy (annual data) France: quarterly data I/1961–IV/1973

Inspection of Table 1 shows that monetary impulses appear significantly in all five countries at the 5 percent level (1.65), and it is significant in four at the 1 percent significance level (2.33). The data offer good reason to reject the chance hypothesis and assign substantial economic significance to the monetary impulse. The fiscal impulse, on the other hand, is not significant in four cases even at very high levels of significance (i e., substantially above 5 percent). Germany yields, however, a different pattern and produces a ∇p statistic well above the 1 percent significance level. The ∇p value for the monetary impulse is, however, almost four times the ∇p value for the fiscal impulse. The external impulses appear significantly (at the 1 percent level) in the case of Germany and Italy. Neither France nor the United States shows any significant positive association between the inflationary accelerations and changes in these impulse measures over the sample period. We find for the Netherlands however a ∇p value for $EI(P)$ almost touching significance level at 1 percent. Remarkable is the irrelevance of the external quantity effect in the Dutch case. There emerges from Table 1 a strong support for the monetary impulse, a comparatively weak case for the role of the fiscal impulse, and some partial support for external impulses based on experiences in Germany, Italy and The Netherlands.

Table 2 summarizes tentative results for three countries bearing on the output-impulse relation. The monetary impulse again is significant for all countries (at the 1 percent level). The fiscal impulse emerges only in the German case and with a fraction of the ∇p value assigned to the monetary impulse. Even the external quantity impulse $EI(Q)$ appears for Germany with much sharper significance than the fiscal impulse. The external quantity impulse does not operate very significantly in the United States and The Netherlands. The external price impulse occurs significantly (at 5 percent) for The Netherlands, and at 1 percent for the United States. The positive association is, of course, defined between output changes and a negative valued $EI(P)$. Tables 1 and 2 convey, in summary, a comparatively strong impression about the importance of the monetary impulse, with a somewhat marginal effect appearing for the fiscal impulse, and a limited effect for the external impulses.

TABLE 2—THE ∇p STATISTIC MEASURING THE ASSOCIATION BETWEEN CHANGING IMPULSES AND ACCELERATION OF OUTPUT

	United States	Netherlands	Germany
FI	9	2	6 6
MI	4 8	4 8	57 9
EI(P)	2 5	1 7	n a
EI(Q)	1 2	1 5	27 0

Note See the explanatory notes to Table 1

The importance of the data in Tables 1 and 2 follows from the circumstance that a comparatively weak test (reflecting comparatively nonconstraining assumptions) is applied to demanding data involving second time differences. It is useful in our judgment to supplement the nonparametric test even after further applications and some technical refinement, with appropriate parametric procedures.

Table 3 reports some regressions from the preliminary work. The regressions involve

either relative changes or accelerations. The results obtained for the Netherlands is indicative of the general pattern. The first regression contains only the four impulses introduced above. The monetary and the external price impulse clearly emerge with leading significance for Dutch inflation. The second regression replaces the domestic monetary growth with a measure of the growth rate of the world money stock. It also includes a measure of "autonomous price changes," i.e., changes in state controlled prices, and an (inverse) index, q, of capacity utilization. The two additional elements have the expected signs and substantial significance, whereas the insertion of the world monetary change lowers understandably the importance of $EI(P)$. The coefficient of determination $\bar{R}^2$ seems quite satisfactory, but the intercept (statistically significant at standard levels) poses a problem requiring attention in further work.

TABLE 3—SOME REGRESSION PATTERNS OBTAINED FOR INFLATION AND OUTPUT FOR SEVERAL COUNTRIES

1. The Netherlands

a. Inflation

$$\hat{p}_{cp} = 2.51 - .13\,EI(Q) + .27\,EI(P) - .01\,FI + .35\,\hat{M}_{-1}$$
$$(2.12)\quad(1.11)\quad(2.02)\quad(-.04)\quad(3.35)$$
$$\bar{R}^2 = .59,\; DW = 1.92$$

$$\hat{p}_{cp} = 2.95 + .01\,EI(Q) + .18\,EI(P) + .03\,FI + .23\,\hat{WM}_{-1} + 1.26\hat{p}_{an} - .44\,q_{-1}$$
$$(3.79)\quad(.13)\quad(1.81)\quad(.22)\quad(2.66)\quad(3.38)\quad(2.77)$$
$$\bar{R}^2 = .84,\; DW = 3.41$$

$$\hat{p}_{cp} = 3.30 + .30\,\hat{WM} + 1.17\,\hat{p}_{an} - .55\,q_{-1}$$
$$(4.83)\quad(3.89)\quad(3.09)\quad(4.40)$$
$$\bar{R}^2 = .82,\; DW = 2.70$$

b. Output

$$\hat{y} = -1.78 + .76\,EI(Q) + .20\,EI(P) - .07\,FI + .05\,\hat{M}_{-1} + .53\,q_{-1}$$
$$(1.23)\quad(5.56)\quad(1.07)\quad(-.30)\quad(.53)\quad(2.32)$$
$$\bar{R}^2 = .80,\; DW = 2.53$$

$$\hat{y} = -.46 + .64\,EI(Q) + .21\,EI(P) + .00\,FI + .25\,\Delta\hat{M}_{-1} + .46\,q_{-1}$$
$$(.35)\quad(5.42)\quad(1.32)\quad(.00)\quad(2.35)\quad(2.46)$$
$$\bar{R}^2 = .86,\; DW = 2.10$$

$$\hat{y} = .46 + .64\,EI(Q) - .06\,\Delta EI(Q) - .03\,\hat{M}_{-1} + .28\,\Delta\hat{M}_{-1} + .31\,q_{-1}$$
$$(.32)\quad(4.58)\quad(.55)\quad(-.36)\quad(2.11)\quad(1.77)$$
$$\bar{R}^2 = .84,\; DW = 1.71$$

Notes: a. variables with a hat sign describe percentage changes
b. regressions are based on annual data
  for inflation 1955–73
  for output 1956–73
c. industrial production is used as output measure
d. $EI(Q)$ measure of external quantity impulse
  $EI(P)$ measure of external price impulse
  money stock is M_1, the narrow measure
  WM measure of world money stock
  q index of capacity utilization (*larger* utilization means *lower* values of q.)
  p_{cp} an index of consumer prices
  p_{an} an index of prices contained in the consumer price (index administered autonomously by the government)
e. numbers in parenthesis below the regression coefficients refer to t-values

2. The United States

a. Inflation

$$\hat{p}_{cp} = 1.15 - .09\,FI + .24\,EI(P) + .31\,\hat{M}_{-1}$$
$$(1.69)\quad(-.54)\quad(1.75)\quad(2.07)$$
$$\bar{R}^2 = .37,\; DW = .91$$

$$\hat{p}_{cp} = .76 + .20\,EI(P) + .36\,\hat{M}_{-1} + .04\,\hat{y}_{-1}$$
$$(1.21)\quad(6.88)\quad(2.87)\quad(.40)$$
$$\bar{R}^2 = .80,\; DW = 1.04$$

b. Output

The dependent variable is $\Delta\hat{y}$, i.e., an acceleration measure of real GNP

$$\Delta\hat{y} = -.17 + .48\,\Delta FI - .09\,\Delta EI(P) + 1.33\,EI(Q) + 1.18\,\Delta\hat{M}$$
$$(-.31)\quad(3.04)\quad(-1.31)\quad(1.76)\quad(5.05)$$
$$\bar{R}^2 = .65,\; DW = 1.91$$

$$\Delta\hat{y} = 2.34 + .24\,\Delta FI - .06\,\Delta EI(P) + .92\,EI(Q) + .88\,\Delta\hat{M} - .68\,\hat{y}_{-1}$$
$$(2.83)\quad(1.73)\quad(-1.08)\quad(1.53)\quad(4.43)\quad(-3.50)$$
$$\bar{R}^2 = .79,\; DW = 1.93$$

$$\Delta \dot{y} = 25.27 + .35\,\Delta FI + 1.02\,\Delta \hat{M} - 30\,cu$$
$$(2.48) \quad (2\,24) \quad (4\,21) \quad (-2\,50)$$

$$\bar{R}^2 = 65,\ DW = 1.72$$

Notes cu index of capacity utilization

The other variables are explained under the regressions describing the Dutch experience.

The sample period covers, 1952 72

3. Italy

$$\hat{p} = -1\,67 + 69\,(\hat{M} - \dot{y}) + .18\hat{i}$$
$$(-1\,90) \quad (6.98) \quad (5.52)$$

$$\bar{R}^2 = 76,\ DW = 2\,39$$

$$\hat{p} = -06 + 52\,\hat{M} - 54\,\dot{y} + 54\,(\hat{p}_{-1} - \hat{p}_{-2})$$
$$(-02) \quad (3\,29) \quad (1\,95) \quad (2\,25)$$

$$R^2 = 46,\ DW = 1\,98$$

$$\hat{p} = 59\,(\hat{W} - \hat{P}r) + 04\,\hat{MP}$$
$$(7\,05) \quad (\,41)$$

the regression was through the origin

$$R^2 = 13,\ DW = 90$$

$$\hat{p} = 2.63 + 32\,(\hat{W} - \hat{P}r) + 12\,\hat{MP}$$
$$(9\,70) \quad (7\,16) \quad (3\,16)$$

$$R^2 = 85,\ D = 1\,77$$

sample mean of dependent variable 4 2

$$\hat{p} = -05\,\hat{S} + 51\,L(\hat{M}) - 06\,cu + 55\,\pi$$
$$(-5\,68) \quad (7\,39) \quad (-5\,76) \quad (7\,45)$$
$$+ 37\,EI(P) - 004\,\dot{y}x$$
$$(4\,59) \quad (\,10)$$

$$\bar{R}^2 = 87,\ DW = 87$$

Notes The first two regressions are based on annual data 1954–73

The last two regressions are based on annual data 1952–73

π anticipated inflation rate based on adaptive scheme

$\dot{y}x$ measure of real growth outside Italy

The term $L(M)$ in the last regression represents a linear combination of lagged values of monetary growth reaching a lag of five

4. Germany

$$\Delta \hat{p} = -39 + 16\,\Delta \hat{M}_{-2} + 09\,\Delta FI_{-1}$$
$$(-1\,13)\,(3\,09) \quad (3\,71)$$
$$+ 21\,EI(P) + 13\,\dot{y}_{-1}$$
$$(4\,11) \quad (2\,11)$$

$$R^2 = 78,\ DW = 1\,48$$

Note This regression is based on annual data covering the period 1958–74

$$\Delta \dot{y} = -36 + 48\,\Delta \hat{M} + 00\,\Delta FI + 65\,\Delta EI(Q)$$
$$(-59) \quad (2\,40) \quad (\,03) \quad (4\,52)$$

$$\bar{R}^2 = .51,\ DW = 1\,91$$

Note This regression is based on annual data covering the period 1956–74

4. France

$$\hat{p} = .20\,\hat{M}_{-2} + .06\,FI + 17\,EI(P) + .10\,EI(Q)$$
$$(8\,88) \quad (1\,12) \quad (4.44) \quad (3.69)$$

$$R^2 = 35,\ DW = 1.2$$

Note Sample period I/1960–IV 1973, quarterly data

There seems to be no problem about the intercept for the United States. The intercept is small relative to the mean value of the dependent variable and the associated significance level is quite high. The monetary and the external price impulse dominate the fiscal impulse. A remarkable improvement of the fit is achieved by including an implicit capacity effect in the form of a lagged value of the relative change in output The capacity effect occurs with the expected sign but with low significance.

The German regressions consider the relation between accelerations, i.e , second time differences. Fiscal acceleration, monetary acceleration and acceleration of external prices all appear to have substantial significance. The regression pattern supports the results of the ∇p tests The German data also suggest that the monetary impulse is subject to the longest lag. We note furthermore the significant occurrence of the capacity effect on the rate of inflation. The intercept is quite low and insignificant in the German case.

The "French regression" exhibits the same long monetary lag already noted in the German case. All impulses with the exception of the fiscal impulse appear with highly significant coefficients But the fit of the regression is comparatively poor (i.e , $\bar{R}^2 = 35$), and the DW statistic poses a question bearing on the interpretation of the positive serial correlation of residuals.

We consider, lastly, the Italian results. The first two regressions exemplify some versions of a quantity theory with nonconstant velocity expressed via the interest change effect $\hat{i}$. The second regression is obtained by replacing i with $r + \pi$ (real rate + anticipated rate of inflation). It is further assumed under the circumstances that r is constant and $\pi = \hat{p}_{t-1}$. It is interesting to note in the second regression

that the coefficients for $\hat{M}$ and $\hat{y}$ are essentially equal in magnitude with opposite signs. We also note an adequate value of the *DW* statistic and a low and nonsignificant intercept. The circumstance is particularly interesting in comparison with the next two regressions representing a "wage-push" mechanism. A regression of the inflation rate on wage changes corrected for productivity changes and import price changes shows very little systematic relation when the regression is forced through the origin. The fit is spectacularly improved when the intercept is estimated freely. But the intercept explains more than 60 percent of the dependent variable's sample mean under the circumstances. The "quantity theory" exhibits on the other hand an essentially vanishing intercept. The "wage-push" formulation clearly obtains no support in comparison with a monetary explanation. The last regression for Italian inflation uses an extended lag pattern for the monetary impulse. This impulse is significantly supplemented with the operation of an external price impulse and the anticipated rate of inflation based on an adaptive process. The fit and the intercept are clearly supportive, but the astonishingly low (compared to the quantity theory regressions) *DW* statistic poses some questions for future examination.

The regressions executed for output yield some complementary results. All three countries examined produce strong patterns with respect to the monetary impulse and the external quantity impulse. It is noteworthy in this respect that in the case of the Netherlands monetary acceleration (i.e., "unanticipated" monetary movements) operate with substantially larger significance on output changes than monetary growth. The acceleration effect of the real variable *EI(Q)* vanishes in comparison to the direct effect of *EI(Q)*. The external price effect and the fiscal impulse are not significant either directly or as accelerations.

The German data exhibit similar results. The monetary impulse and the external quantity impulse *EI(Q)* dominate the regression. The fiscal impulse vanishes in significance with respect to output (in contrast to inflation). The reader should note, when judging the coefficient of determination, that the regressions involve accelerations of the underlying variables. The *DW* statistic is also adequate and the intercept is small and not significantly different from zero.

The *U.S.* regressions use acceleration of the impulses and output. A somewhat different pattern emerges in this case, however. The coefficient of determination is high for the kind of variables used, the *DW* statistic adequate, and the intercept of the first regression (on impulse variables only) small and not significant. But the fiscal impulse dominates the external quantity impulse and joins the monetary impulse as a significant force operating on output. The other regressions for *U.S.* output insert additional capacity effects. This insertion substantially modifies the position of the intercept and lowers the coefficient and significance of the fiscal impulse. The smallest change is in the response to monetary acceleration.

III. Conclusion

The results obtained to date from the group's empirical investigation yield a strong case for the operation of financial impulses, and particularly for the operation of monetary impulses. There is also some evidence, occasionally strong, of the operation of external impulses on inflation and output changes (or accelerations).

The sustained and accelerating inflation emerging since the middle 1960's evidently results from gradual shifts in budgetary and monetary policies of Western countries. The reason for this development lies beyond the scope of this limited report. An exploration of the causes at work moves us into an analysis of "political and sociological facts" which offer a fruitful and exciting area for extension of economic analysis.

Export Prices and the Transmission of Inflation

By IRVING B. KRAVIS AND ROBERT E. LIPSEY*

We report here on studies of price behavior that reveal a very different world from that assumed in most models of the transmission of price movements. We find that there are sometimes substantial and prolonged divergences between the export price movements of different countries for the same or closely related products and notable differences within countries between export and domestic price changes. We summarize some of the evidence from our earlier studies[1] and extend it to a wider range of commodities and to recent years, utilizing new price indexes presented here for the first time.

The prevailing view of the international price system, on the other hand, is based on the "law of one price." In terms of absolute levels of prices, the prices of identical internationally traded goods are held to be the same everywhere, after making due allowance for transfer costs. Price changes for exports and tradeables generally are regarded as being kept closely aligned by the possibility of commodity arbitrage and by the possibility of substitutions in production and consumption. This view has been most explicitly adopted by the monetarists and even extended, to varying degrees by different members of the school, to nontraded goods as well.[2]

If, as we seek to demonstrate, the export and domestic prices of the same product in a given country and the export prices of the same product from different countries can differ and can move differently, the links between the price systems of different countries will be looser. If export and domestic prices can differ, the impact of an inflating country's increased demand for imports may be expected to be smaller on domestic prices of other countries than on their export prices; traded goods, though serving as a transmission channel for inflation, will not necessarily transmit the full amount of inflation. Thus, real changes in quantities produced and traded will be more affected by changing relative prices than in a world in which the law of one price prevails, and the scope for independent monetary and fiscal policies will be somewhat greater.

This real world price behavior reflects both static and dynamic factors that interfere with tendencies toward a one price world. From a static standpoint, many firms involved in international trade, particularly in manufactures, are in the position of a discriminating monopolist faced with separate markets, each characterized by a different demand elasticity. If we make the usual assumption that the monopolist has short-run rising marginal costs, an expansionary

*Professors of Economics, University of Pennsylvania, and Queens College, City University of New York, respectively, and senior research staff members of the National Bureau of Economic Research *(NBER)*

The basic data collection and construction of price indexes for this paper were done under several grants to the *NBER* from the National Science Foundation and extended to recent years under a contract with the Office of Competitive Assessment of the U.S. Department of Commerce. The views reported here do not necessarily reflect those of either agency. This report has not undergone the review accorded official *NBER* publications, in particular, it has not been submitted for approval by the Board of Directors and therefore is not a publication of the National Bureau.

We are indebted to Mary Boger, Daniel Gottlieb, and Judy Rosenzweig for assistance in the preparation of the paper and to Eliot Kalter of the University of Pennsylvania for the matching of *U.S.* export and domestic price data for the latter part of the period. We thank Arthur Bloomfield, Alan Heston and Richard Marston for helpful comments on an earlier version of this paper. We also benefited from comments made at a presentation at the International Economics Workshop at the University of Pennsylvania.

[1]See Kravis and Lipsey (1971 and 1974). More recently Peter Isard has called attention to international price differences.

[2]See Marina von N. Whitman and Jacob Frenkel and Harry Johnson.

stimulus in a foreign country should lead him to raise his export price more than his domestic price. The higher export price provides an inducement to shift sales away from the home market to foreign markets and reduced quantities at home eventually lead to price increases. Thus the model of the price discriminator leads to differential changes in price relationships and to shifts in quantities traded.

The applicability of this or any other static equilibrium model is complicated by the dynamic forces continually operating in international trade. There have been shifts in comparative advantage which have led to changes in export shares; for example, Japan's share in "world" manufactured exports rose by more than 70 percent during the 1960's while the *U.K.* share dropped by more than a fourth and that of the United States by more than a tenth.[3] However, the adjustment to any given change in comparative advantage is not instantaneous but gradual. Lack of knowledge, uncertainty regarding the reliability of new suppliers, the reluctance to give up a satisfactory relationship with customary suppliers and commitments to a given type of equipment because of previous purchases or stocks of spare parts may all explain the failure of buyers to respond immediately to price differences. They may explain too why it is sometimes necessary for substantial price differences to exist and to persist over protracted periods for sellers to overcome the inertia of buyers in patronizing customary sources. Selling at a low price is, after all, the traditional way of breaking into a market and expanding market shares.

Given this freedom of the export prices of a given country to deviate from the domestic prices of the same goods and from the export prices of other countries for the same goods for considerable periods of time, we may expect something like the following series of events in an inflation: 1) an inflation will raise a country's domestic prices more than its export prices; 2) the rise in the inflating country's prices will tend to pull up the export prices of other countries, in absolute terms and relative to their domestic prices; 3) eventually the rise in export prices leads to at least some increase in domestic prices as well; 4) the decrease in the ratio of export to domestic prices in the inflating country will cause a shift from export to domestic sales, and the opposite change in the export/domestic price ratio in the other countries will bring about a shift from domestic to export sales; the rise in the inflating country's export prices relative to other countries' export and domestic prices will reduce its export share in world markets. Thus changes in export shares may be attributable both to supply side shifts, stimulated by changes in incentives to export, and the demand side responses to changes in relative export prices between the inflating country and the other countries. We expect at least some of these changes to occur slowly, perhaps over a two- or three-year period.

In the equations described below we test several of the links in these price and quantity sequences. Because our purpose is to call attention to overlooked links in the transmission mechanism, involving discriminatory pricing of exports, we do not try in this paper to test every step in the sequence, nor do we deal with the macroeconomic implications of the price and quantity changes we discuss.

I. Prices May Differ Substantially for Competitive Products Exported by Different Countries

Documentation of the existence of substantial differences in the export prices of different countries may be found in a *NBER* study by the present authors dealing with international price competitiveness for metals, metal products, machinery, and transport equipment.[4] While some differences were found in all 6 of the 2-digit Standard International Trade Classification *(SITC)*[5] categories included in the study, the largest differences were in iron and steel (*SITC* Division 67). Japanese prices averaged 30 percent less than those of the United States, German prices 24 percent less and the *U.K.* prices 22 percent less. Differences in individual

[3]The share comparisons are for the years 1960 and 1970, the "world" consists of the 14 major industrial countries. See U.S. Department of Commerce

[4]Kravis and Lipsey (1971)

[5]*Standard International Trade Classification, Revised*, Statistical Papers, Series M, No. 34 (New York: United Nations, 1961).

categories were as large as 43 percent for Japan in the case of iron and steel wire (*SITC* 677) and 40 percent for Germany in the case of bars and rods (*SITC* 673.2) and tube and pipe fittings (*SITC* 678.5). These differences persisted more or less over the entire period covered by the study, 1953–64. The period was one in which the *U.S.* share in the iron and steel exports of the 21 Organization for Economic Co-Operation and Development countries declined from 19 percent to 10 percent and that of the *U.K.* from 14 percent to 9 percent, while the German share rose from 12 to 18 percent and the Japanese share from 5 to 14 percent. Similar, though less dramatic differences in prices and changes in shares were found in nonelectrical machinery and electrical machinery. For this period, at least, notable and even substantial price differences persisted while the low price sellers gradually expanded their market shares and the high priced sellers saw their shares contract.

These findings suggest that price differences like these may accompany other shifts in trade shares in individual product classes. Since these shifts are continually occurring, such disequilibrium situations in which markets have not fully adjusted to changes in comparative advantage may be the norm rather than the exception.

II. Export Prices are More Variable and More Sensitive to Foreign Events than are Domestic Prices

Utilizing the time series data on matching export and wholesale prices for 1953–64, assembled in the earlier study, and official *U.S.* and German price series, we are able to compare the movements of export and domestic prices in the United States and Germany. The official series were used to interpolate between the years 1953, 1957, and 1961–64 for which the earlier study gathered a substantial body of primary data, and to extend the series from 1964 to 1975. In the case of the United States we took advantage of the important new export price series prepared by the Bureau of Labor Statistics *(BLS)* dating from 1964. Owing to the limited commodity coverage of both our own earlier work and the more recent *BLS* work, the *U.S.* coverage in the following analyses is confined to machinery and equipment (*SITC* 7); that of Germany includes not only machinery and equipment but all of manufacturing. The domestic price indexes are built up by assigning the wholesale price series for individual commodities to the appropriate 4-digit *SITC* category, by getting an index for each category by taking unweighted averages of the component item indexes, and by aggregating the 4-digit categories with weights consisting of the 1963 exports of each country. The export price indexes and ratios of export to domestic price changes were prepared in an analogous way. The series are presented in Table 1.

While changes in export and domestic prices are closely related, export prices fluctuate more widely both for *U.S.* machinery and transport equipment (*SITC* 7) and for German manufacturing as a whole (*SITC* 5–8). This is indicated by the fact that the coefficient for domestic price is always above one in bivariate equations when the export price is the dependent variable, but the coefficient for export price is below one in equations when the domestic price is the dependent variable. In the case of *U.S.* machinery and transport equipment, for example, the coefficient for domestic price in the export price equation was 1.32 while the coefficient for export price in the domestic price equation was .69.

For both *U.S.* machinery and transport equipment and German manufactures as a whole, changes in foreign prices affect both export and domestic prices, as can be seen in equations (1) through (4).[6] We found little relation between exchange rate changes and prices except for one unexpected negative coefficient for the lagged exchange rate change on German domestic prices. The effect of foreign price changes on both *U.S.* and German prices was substantially larger for export than for domestic prices, confirming the second step of the price sequence described above.

[6]These equations, and all others in the paper, are arithmetic, with variables in the form $\frac{t_1}{t_0} \cdot 100$. In equations (1) through (4) *U.S.* and German unit labor costs in manufacturing are included to reflect changes in domestic cost conditions which affect both export and domestic prices.

TABLE 1—U.S., U.K., AND GERMAN DOMESTIC PRICES, U.S AND GERMAN EXPORT PRICES, U.S AND GERMAN EXPORT/DOMESTIC PRICE RATIOS
(1963 = 100)

	SITC 7							SITC 5-8	
	Domestic Prices			Export Prices		Export/Domestic Prices		Export Price	Export/ Domestic Price
	U.S ($)[a] (1)	U.K. (£) (2)	Germany (DM) (3)	U.S ($)[a] (4)	Germany (DM) (5)	U.S ($) (6)	Germany (DM) (7)	Germany (DM) (8)	Germany (DM) (9)
1953	85.7	NA	87.9	85.4		99.6			
1954	85.9	82.1	85.9	85.3		99.3		91.7	103.9
1955	86.0	84.5	85.9	85.2		99.1		93.9	105.5
1956	89.6	88.1	87.8	88.1		98.3		95.9	105.8
1957	93.8	91.7	90.3	92.3		98.4		97.6	105.1
1958	95.6	94.0	91.2	94.0	94.0	98.3	103.0	98.1	104.6
1959	97.9	94.6	90.5	96.0	94.1	98.1	103.9	97.4	104.4
1960	98.5	95.2	91.9	97.4	95.4	98.9	103.4	98.8	104.8
1961	98.6	97.4	94.4	97.7	97.4	99.1	102.6	99.5	102.7
1962	99.8	98.7	99.3	99.2	99.2	99.4	100.0	100.0	100.6
1963	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1964	101.5	101.9	101.6	101.4	101.4	99.9	99.7	102.4	100.8
1965	103.5	104.7	104.5	102.6	104.4	99.1	99.9	105.1	100.4
1966	107.7	106.7	107.7	103.8	107.2	96.4	99.5	107.3	99.7
1967	110.8	108.0	107.6	107.3	108.1	96.8	100.5	107.3	100.3
1968	114.8	111.8	107.5	110.9	107.7	96.6	100.1	106.2	99.2
1969	118.7	114.8	109.3	115.2	109.1[b]	97.1	99.5[b]	107.2[b]	99.4
1970	124.3	124.0	117.9	120.4	117.6	96.9	99.5	115.6	100.2
1971	129.9	135.7	127.0	125.4	125.7	96.5	99.1	120.6	98.9
1972	133.2	144.1	133.0	128.9	130.7	96.8	98.3	123.8	97.8
1973	137.1	153.4	139.8	134.2	137.4	97.9	98.4	132.8	98.4
1974	154.0	184.4	153.9	154.4	149.1	100.3	97.0	155.3	98.9
1975	170.6	NA	NA	180.4		105.7			

Sources: Col. (1) *BLS* Wholesale Price Data for individual items in 4-digit SITC subgroups matching Col. (4), with *U.S.* export weights.

Col. (2) Calculated by *U.K.* Board of Trade (later Division of Trade and Industry) from detailed price data with *U.K.* export weights for 1972-74, *OECD* export weights for 1953-71.

Col. (3) Detailed German price series aggregated by *NBER* with German export weights.

Cols. (4) and (5) Detailed export price series from Kravis and Lipsey (1971) and *BLS (U.S.)* and Federal Statistical Office (German) weighted by *NBER* with *U.S.* and German export weights.

Cols. (6) and (7) Ratios of detailed export to domestic price series, weighted by *NBER* with *U.S.* and German export weights.

Col. (8). Calculated from detailed German export price series, from Kravis and Lipsey and Federal Statistical Office, aggregated by *NBER* with German export weights.

Col. (9) Calculated from export/domestic price ratios, for 4-digit *SITC* subgroups and aggregated by *NBER* with German export weights.

[a]Excluding *SITC* 722.1 and 729.3

[b]Three percent export tax removed from price

$$(1)\quad EXPUS = \underset{(2.92)}{46.3} + \underset{(2.44)}{.286}\,ULCUS - \underset{(1.22)}{.099}\,XR\$/£ + \underset{(4.04)}{.460}\,DOMUK + \underset{(1.67)}{.127}\,XR\$/£(-1)$$

(2) $DOMUS = 53.7 + .263\, ULCUS$
$(2.90)\ (1.92)$
$+ .336\, DOMUK$
(2.52)
$+ .034\, DOMUK(-1)$
$(.21)$
$- .074\, XR\$/£$
$(.78)$
$- .089\, XR\$/£(-1)$
(1.00)
$\bar{R}^2 = .86 \qquad DW = 1.63$

(3) $EXPGE = -14.7 + .423\, ULCGE$
$(.28)\ (2.82)$
$+ .710\, DOMUS$
(2.56)
$- .054\, DOMUS(-1)$
$(.12)$
$+ .071\, XRDM/R$
$(.67)$
$- .012\, XRDM/\$(-1)$
$(.05)$
$\bar{R}^2 = .88 \qquad DW = 1.83$

(4) $DOMGE = 55.1 + .502\, ULCGE$
$(1.84)\ (5.77)$
$+ .356\, DOMUS$
(2.21)
$- .222\, DOMUS(-1)$
$(.82)$
$+ .093\, XRDM/\$$
(1.49)
$- .279\, XRDM/\$(-1)$
(2.18)
$\bar{R}^2 = .94 \qquad DW = 1.82$

EXPUS, *EXPGE*, *DOMUS*, and *DOMGE* = *U.S.* and German export and domestic price indexes

ULCUS and *ULCGE* = *U.S.* and German unit labor costs in manufacturing.

DOMUK = *U.K.* domestic price index.

XR\$/£ = *U.S.* exchange rate in \$ per £.

XRDM/\$ = German exchange rate in *DM* per \$.

Equations (1) and (2) refer to *SITC* 7; equations (3) and (4) to all manufacturing.

All equations refer to the period 1953, 1954, or 1955 to 1974. Figures in parentheses are t = ratios.

III. There is Some Evidence that the Export/Domestic Shipment Ratio and the Export/Domestic Price Ratio Move Together

As is implied by the stronger impact of foreign price changes on export prices than on domestic prices, the ratio of export to domestic prices changed over time in both the United States and Germany. The range of variation in the ratio was 9.7 percent for *U.S. SITC* 7, 7.0 percent for German *SITC* 7, and 8.1 percent for German manufacturing as a whole.

Changes of less than 10 percent in the export/domestic price ratio over a twenty-year period may not seem to be very large, but when account is taken of profits/sales ratios (about 4 percent in 1970 for the *U.S.* corporations roughly approximating *SITC* 7[7]), such swings in the relative prices of export and domestic sales imply large shifts in the relative profitability of exports and domestic sales. We should expect to see at least the more notable changes in the export/domestic price ratio associated with corresponding changes in exports relative to domestic shipments.

Both the *U.S.* and German records give some support to these expectations. For *U.S.* machinery and equipment, the export/domestic price ratio (1963 = 100) fluctuated within a narrow range (96.4 to 97.7) between 1953 and 1972, and there is little evidence of any association between this ratio and the export/shipment ratio.[8] Beginning in 1972, however, the export/domestic price ratio rose sharply (to 105.7 in 1975), and the export/shipment ratio also increased to new levels (from 9.7 percent in 1972

[7]The 4 percent refers to the ratio of net income before tax to sales, after tax income was less than 2 percent of sales (*Statistics of Income, 1970 Corporate Income Tax Returns*, p. 18, Industries 25 through 28)

[8]Exports and shipments of machinery and transport equipment from various issues of the *Survey of Current Business*, *Foreign Commerce and Navigation of the U.S.: 1965* (Bureau of the Census), and U.N. *Commodity Trade Statistics*

to 17.2 percent in 1975). Perhaps sellers require large changes which seem likely to persist before they are led to reorient their marketing activities.

The German data for all manufacturing are dominated by trend. However, when the export/domestic price ratio was above its trend the export/domestic shipments ratio also tended to be above its trend, and the years when both were below the trend also tended to coincide (see Figure 1).[9] The relationship between the deviations of the series from their straight line trends is as follows):

$$(5) \qquad X/O = .05 + 1.52 P_X/P_D \quad \bar{r}^2 = .30 \quad DW = 1.27$$
$$(0.09)(3.09)$$

where X/O is the deviation from trend of the ratio of manufactured exports to manufactures output and P_X/P_D is the deviation from trend of the export/domestic price ratio. We thus have some confirmation of step 4 of the sequence described earlier.

IV. Changes in the Export/Domestic Price Ratio Appear to be Related to Changes in Exchange Rates and Foreign Prices

A useful working hypothesis about the reasons for the changes in the export/domestic price ratio and in the export/shipments ratio is that they are to be found in the differences between domestic cyclical conditions and those prevailing abroad. In addition, there will be secular influences affecting particular products or product sectors as comparative advantages change; an industry's export/domestic price ratio may, for example, decline during the period of rapid expansion in foreign markets.

We use relative (foreign to home) price movements as a means of summarizing relative cyclical and secular conditions. For this purpose we compare foreign countries' domestic prices with $U.S.$ domestic prices for the same groups of products, and we refer to the resulting ratio as the "relative" rate of inflation. The rationale is that we regard the changes in domestic prices as measures of the pressures of demand against resources, which we can use to test our hypothesis about the nature of the causes of changes in the export/domestic price ratio.

Expansion abroad relative to that at home should raise the home country's export/domestic price ratio with varying lags depending upon the extent of unused capacity at the start. Relative expansion at home should have the opposite effects.

Although the relative rates of inflation measure the relative pressures on resources at home and abroad, the impact of these relative pressures on the home country's export/domestic price ratio cannot be fully assessed unless account is taken also of changes in exchange rates. A depreciation of the home currency, for example, should have the same effect on the ratio as a uniform increase in all foreign prices relative to home prices.

Equations (6) and (7) explain changes in the $U.S.$ export/domestic price ratio for machinery and transport equipment by changes in foreign rates of inflation relative to $U.S.$ rates and by changes in $U.S.$ exchange rates relative to each currency.[10] Since the exchange rate variable

$$(6) \qquad \frac{EXPUS}{DOMUS} = \underset{(10.01)}{93.7} + \underset{(.14)}{.015}\frac{DOMGE}{DOMUS} - \underset{(1.04)}{.111}\frac{DOMGE(-1)}{DOMUS(-1)} + \underset{(.30)}{.014}\,XR\$/DM + \underset{(3.07)}{.142}\,XR\$/DM(-1)$$
$$\bar{R}^2 = .32 \qquad DW = 1.80$$

[9]Exports from various issues of *Statistisches Jahrbuch fur die Bundesrepublik Deutschland* and *Wirtschaft und Statistik*, Statistisches Bundesamt, Wiesbaden, manufacturing output from various issues of U.N. *Yearbook of National Accounts Statistics*

[10]Since the $U.S.$ domestic price index is in the denominator of both the export/domestic price ratio and the relative rate of inflation, the coefficient of the latter may be positively biased if there are errors in the $U.S.$ domestic price measure. It is clear, however, from equations (1) through (4), relating export prices to foreign prices and domestic prices to foreign prices, that any such bias is not the main reason for the positive coefficients in equations (6) and (7).

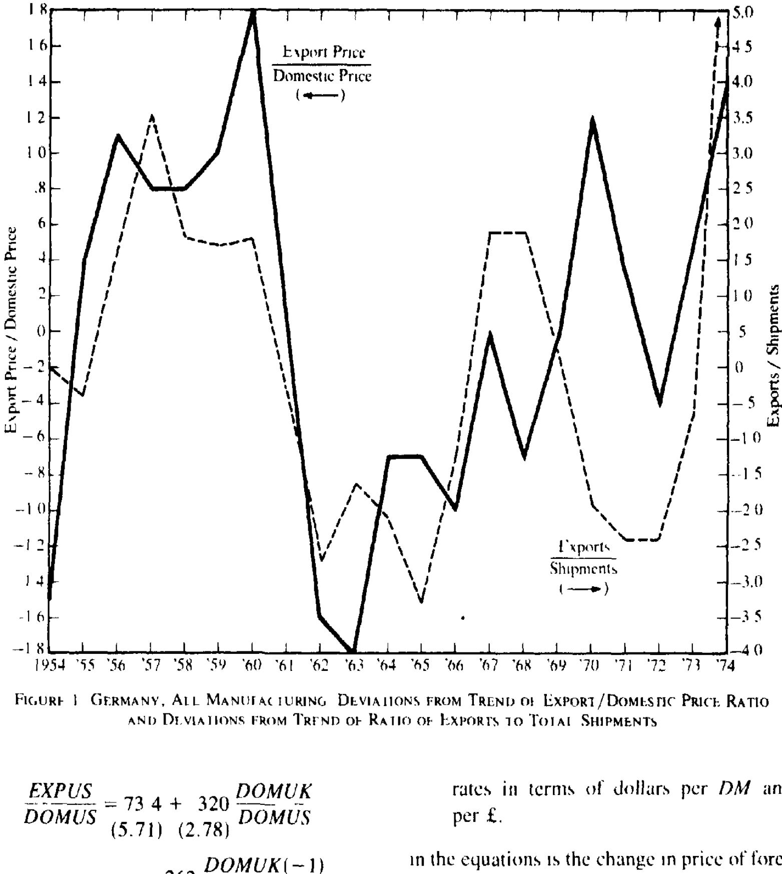


FIGURE 1 GERMANY, ALL MANUFACTURING DEVIATIONS FROM TREND OF EXPORT/DOMESTIC PRICE RATIO AND DEVIATIONS FROM TREND OF RATIO OF EXPORTS TO TOTAL SHIPMENTS

$$(7)\quad \frac{EXPUS}{DOMUS} = \underset{(5.71)}{.734} + \underset{(2.78)}{.320}\frac{DOMUK}{DOMUS} - \underset{(1.60)}{.262}\frac{DOMUK(-1)}{DOMUS(-1)} + \underset{(1.71)}{.226}\frac{DOMUK(-2)}{DOMUS(-2)} - \underset{(.15)}{.010}\, XR\$/£ - \underset{(1.64)}{.108}\, XR\$/£(-1) + \underset{(1.37)}{.098}\, XR\$/£(-2)$$

$$\bar{R}^2 = .40 \qquad DW = 1.76$$

$XR\$/DM$ and $XR\$/£$ = *U.S.* exchange rates in terms of dollars per *DM* and per £.

in the equations is the change in price of foreign currency, the effect of a *U.S.* devaluation is in the same direction as that of a rise in foreign prices. Our expectations therefore are that the coefficients of both independent variables will be positive. It can be seen that the German influence on the *U.S.* export/domestic price ratio was mainly through exchange rate changes; indeed, the only significant coefficient is that for the exchange rate with a one year lag. The *U.K.* influence, on the other hand, was through the relative rate of inflation, most strongly on a current basis. These differences call attention to an important difference in the economic history of the *U.K.* and Germany during this period which is summarized in the following figures:

	1969/55			1974/69			1974/55		
	Germany	United Kingdom	United States	Germany	United Kingdom	United States	Germany	United Kingdom	United States
Industrial prices (own currency)	120.8	136 4	122 6	144.2	163 4	145 1	174.2	222 8	177 9
Exchange rate ($/foreign currency)	107 0	85 6	—	151 7	97 9	—	162 3	83 8	—
Industrial prices ($)	129 3	116 8	122 6	218 8	160 0	145 1	282 7	196 9	177 9

German and *U.S.* rates of inflation were very similar, but there was a large rise in the price of the *DM*, which meant that the exchange rate change accounted for all of the rise in German prices in dollars relative to *U.S.* prices. *U.K.* inflation was more rapid than that of either the United States or Germany, but the price of the £ fell, largely offsetting the rapid inflation with respect to the United States at least. Thus in comparing the *U.K.* and the United States we have large and offsetting changes in relative inflation and exchange rates, while the comparison between Germany and the United States is entirely a story of exchange rate changes. That is probably the explanation for the lack of significant German price coefficients in the *U.S.* equation and the significance of the *U.K* price coefficients. Also it must be borne in mind in assessing the relatively low explanatory power of the equations that each seeks to explain the change in the *U S* export/domestic price ratio in terms of the events in a single foreign country, whereas it is the events in a large number of other countries that are involved. The same point applies to the equations for the German export/domestic price ratio presented below.

Similar relations for the German export/domestic price ratio for all manufactures vis-à-vis *U.S.* and French inflation rates and exchange rates are shown in equations (8) and (9).[11]

[11]Since for Germany we have the export/domestic price ratio for all manufacturing, we have simply used the wholesale industrial price indexes of France and the United States vis-à-vis that of Germany to form measures of relative inflation. A similar equation using *U K* prices showed no significant relationship.

$$(8)\quad \frac{EXPGE}{DOMGE} = \underset{(7.47)}{72.8} + \underset{(3.16)}{.276}\,\frac{DOMUS(-1)}{DOMGE(-1)} - \underset{(.17)}{.007}\,XRDM/\$(-1)$$

$$\bar{R}^2 = 31 \qquad DW = 1.77$$

$$(9)\quad \frac{EXPGE}{DOMGE} = \underset{(5.69)}{101.3} + \underset{(1.11)}{.052}\,\frac{DOMFr}{DOMGE} + \underset{(1\ 71)}{.098}\,\frac{DOMFr(-1)}{DOMGE(-1)} - \underset{(2.67)}{.179}\,\frac{DOMFr(-2)}{DOMFr(-2)} + \underset{(.92)}{045}\,XRDM/Fr - \underset{(1.09)}{.054}\,XRDM/Fr(-1) + \underset{(.52)}{021}\,XRDM/Fr(-2)$$

$$\bar{R}^2 = .51 \qquad DW = 2.26$$

EXPGE/DOMGE = German export/domestic price ratio
DOMUS/DOMGE, *DOMFr/DOMGE* = *U.S.* and French rates of inflation relative to that of Germany
XRDM/\$, *XRDM/Fr* = German exchange rates in terms of *DM* per $ and French franc

The relative rates of inflation have the larger impact in both cases, with a one-year lag for the United States, and with a mixture of current and lagged effects for France. There is a hint (in the French equation and in unpublished

U.S. equations)—in the form of the negative coefficients for the relative rate of inflation lagged two years—of what we would expect a priori, namely that the change in the export/domestic price ratio is eventually likely to reverse itself.[12]

These equations relating foreign price changes to the export/domestic price ratio reinforce the earlier finding that the effects are stronger on export than on domestic prices.

V. Conclusions

In general, we find that commodity markets for manufactured goods are sufficiently tied together that a rise in an important country's price level lifts foreign prices, sometimes immediately and sometimes after a year or so. However, there is no one to one correspondence of price changes between the major industrial countries. Furthermore, export prices can and do differ from domestic prices of the same or closely competitive goods. A consequence is that a country's response to foreign price changes, if they are large and persistent or if trade is very important, includes not only effects on export and domestic prices but a stronger impact on export prices. The variability of the export/domestic price ratio provides an incentive for a shift between export and domestic sales and adds an element of elasticity to the supply curve for exports which cushions the impact on domestic prices of foreign inflation. An implication is that an individual country can have domestic price increases which are not fully reflected in its export prices, that the export prices of the inflating country need not be matched by or restricted to the increased export prices of other countries, and that any consequent increase in export prices in other countries need not lead to an equivalent increase in their domestic prices. A more general inference of the resulting flexibility in the links between national price systems is that a degree of freedom exists in domestic economic policies in an interdependent world that has not generally been taken into account.

[12]This assumes that the long-run elasticities of both export supply and domestic supply are greater than the short-run elasticities. Thus either export prices will eventually move towards domestic prices, or domestic prices towards export prices, or both.

REFERENCES

Jacob A. Frenkel and Harry G. Johnson (eds.), *The Monetary Approach to the Balance of Payments*, Toronto and Buffalo 1976.

Peter Isard, "How Far Can We Push the Law of One Price?" International Finance Discussion Papers, No. 84, Federal Reserve Board, Washington, May 1976.

Irving B. Kravis and Robert E. Lipsey, *Price Competitiveness in World Trade*, New York 1971, 263–68.

———, "International Prices and Price Proxies," *The Role of the Computer in Economic and Social Research in Latin America*, New York 1974.

Marina von N. Whitman, "Global Monetarism and the Monetary Approach to the Balance of Payments," *Brookings Papers*, *3*, 1975.

United Nations, *Standard International Trade Classifications, Revised*, Statistical Papers, Series M, No. 34, New York 1961.

U.S. Dept. of Commerce, *International Economic Indicators and Competitive Trends*, Washington, June 1976, 57.

U.S. Dept. of the Treasury, Internal Revenue Service, *Statistics of Income, 1970: Corporate Income Tax Returns*.

A "Monetarist" Analysis of the Generation and Transmission of World Inflation: 1958–71

By MICHAEL PARKIN*

A widely accepted definition of inflation is that it "is a process of continuously rising prices, or equivalently, of a continuously falling value of money" (David Laidler and Michael Parkin, p. 741). An important observation suggested by this definition is that, in a world with *one* money, or equivalently with many monies linked to a single monetary standard via fixed exchange rates, there is *one* rate of inflation. Of course, index numbers may be computed for subaggregates of goods and services, some of which refer to particular geographical areas—countries—but rates of change in these indices do not measure inflation. Rather they measure a mixture of inflation and relative price changes. This observation has important implications for the analysis and explanation of inflation during the last two decades, for the period from the middle 1950's to 1971 was characterized by a single monetary standard[1] and hence only one inflation rate has to be explained. This paper attempts to explain that inflation rate. It contains no new theoretical ideas and no new empirical results. Rather it encapsulates in a short space the key results which have emerged from the by now large volume of work on the explanation of inflation in the fixed exchange rate world and attempts to identify the major questions which head the agenda of future research.

I. World Average Inflation in the Fixed Exchange Rate World: 1958–71

In a world[2] with one monetary standard and one rate of inflation the explanation of inflation must be sought at the level of the world economy rather than at the national level. To study inflation at the world aggregate level is to follow a tradition begun by Jean Bodin, and David Hume and recently popularized by Robert A. Mundell and Harry G. Johnson (1972).[3] However, it in no way forces onto the analysis a monocausal "monetarist" explanation of inflation as presented in those earlier world aggregate analyses. On the contrary, it provides the simplest possible environment, that of a closed economy, for developing an entirely eclectic framework within which to discriminate among competing hypotheses on the causes of inflation.

Such a framework was developed in Malcolm R. Gray and Michael Parkin. The framework combines three interacting propositions which are capable of embracing all views on the generation of inflation. First, the rate of inflation is influenced by inflation expectations, excess demand and a variety of cost-push factors (including direct wage and price controls); secondly, inflation expectations respond to the history of inflation and to expectations of the

*University of Western Ontario. I am grateful to George Zis for countless hours of discussion on the subject of this paper over several years and to Robin Bade, Peter Howitt, and David Laidler for comments on an earlier draft.

[1]There were only six changes in exchange rates among the major currencies between 1956 and 1976; the revaluations of the D-Mark in 1961 and 1969 and the Dutch Guilder in 1961, and the devaluations of the French Franc in 1958 and 1969 and of Sterling in 1967.

[2]The "world" here is the aggregate of countries which (ignoring the minor exceptions noted above) maintained a fixed exchange rate with the *U.S.* dollar and full convertibility. The "world" as used in much of the empirical work to be reported below is the "Group of Ten," i.e., Belgium, Canada, France, Germany, Italy, Japan, The Netherlands, Sweden, United Kingdom, and United States.

[3]Recent contributions which have also taken a world aggregate view are: Johnson (1975), Arthur B. Laffer and David I. Meiselman, Laidler and A. R. Nobay, Parkin and George Zis, eds. (1976 a,b), Edward S. Shaw, Alexander K. Swoboda (1975b), Ronald L. Tiegen, and H. Joannes Witteveen. For an excellent doctrinal history on the approach see Jacob Frenkel.

movements of the exogenous variables which are believed to cause inflation; thirdly, excess demand is determined by the money supply, fiscal policy, and the behavior of the actual and expected price level. This eclectic view specializes to a variety of extreme positions by assuming some of the potential interlinkages to be either weak or absent. Two such extremes are worth singling out: "cost-push" and "monetarist."

The "cost-push" extreme denies the connection between excess demand and the rate of inflation but accepts the other propositions. It also adds a fourth proposition namely that the money supply is endogenous and passively responds to movements in prices and aggregate excess demand. Thus, on this extreme view, inflation is caused by cost-push factors while a passive monetary policy ensures that real output does not decline (or not often) so that the push factors come through in both the inflation rate and the monetary expansion rate. Further, on this view, the way to control inflation is to attack it at its supposed source with direct controls on wages and prices, preferably in an internationally synchronous fashion.

The "monetarist"[4] extreme view emphasizes the role of excess demand and inflation expectations (the latter with a coefficient of unity) as the sole proximate determinants of systematic movements in the rate of inflation with push factors affecting (at most) the detailed timing of random (zero mean) movements in its rate. Inflation expectations, however formed, will, in a steady state, line up with the actual rate of inflation. Finally, excess demand depends only on the behavior of the money stock, fiscal policy being unimportant, and further, aggregate excess demand is homogeneous of degree zero in the money stock and actual and expected prices. The policy implication of this view is that control of the money supply is both necessary and sufficient for the control of inflation and that the abandonment of money supply control in favor of aiming for a target real output level, other than zero excess demand, will lead to explosive price level and money supply behavior.

Empirical work on the world aggregate (Group of Ten—G-10) economy enables a start to be made in discriminating between these two extremes (and amongst the many intermediate eclectic positions embodied in the three general propositions) stated above. On price setting and expectations formation at the G-10 level, Nigel Duck, et al. estimated equations of the standard form:

$$\Delta p = \alpha x_{-\tau} + \delta \Delta p^{e} + u \tag{1}$$

where p is the price level (in natural logarithms), x is proportionate excess demand (measured as the deviation from trend of the logarithm of real output), τ denotes a time lag, e denotes expectation, Δ is the first difference operator, u is an error term, and α and δ are positive parameters.

If the estimated value of α in equation (1) is not significantly different from zero and if there are large systematic errors in prediction (large variance and temporal dependence in u), then there is a presumption in favor of the "cost-push" extreme. If, alternatively, α is significantly nonzero, δ is not significantly different from unity and u does not display systematic autocorrelation, then the "monetarist" extreme is not rejected. Using a Box-Jenkins *ARIMA* procedure to estimate the forecasting scheme to generate Δp^{e}, equation (4) below, Duck et al. report the following results using quarterly data, 1956–71 for G-10 (t–statistics in parentheses):[5]

[4] I do not want to devote any space to defending this definition of "monetarist." For an extensive discussion of this and related issues, see Jerome L. Stein.

[5] For full details of data sources and methods, and estimation procedures, see Duck et al. (1976). The results reported here are not the best fitting but the simplest reported by Duck et al. A more complex expectations scheme than that embodied in equation (4) gave even better results in the price equation.

(2) $\Delta p = 0.500 + 0.204x_{-1} + 0.814\ \Delta p^e$
$\qquad [1.23] \quad [3.58] \qquad [6.88]$
$\bar{R}^2 = 0.533 \quad D.W. = 2.144$

(3) $\Delta p = -0.100 + 0.184x_{-1} + \Delta p^e$
$\qquad [0.72] \quad [3.31]$
$\bar{R}^2 = 0.513 \quad D.W. = 2.188$
$F = 2.465$

(4) $\Delta p^e = 0.309\ \Delta p_{-1} + 0.691\ \Delta p^e_{-1}$
$\qquad [3.22] \qquad [7.19]$
$\bar{R}^2 = 0.378$

Equation (3) imposes the restriction that the coefficient on Δp^e is unity and the reported F is that associated with the test of the hypothesis that the restriction is true. That restriction cannot be rejected at the 5 percent level, nor can the hypothesis of no first order autocorrelation; nor further can the hypothesis that the coefficient on x_{-1} is nonzero. Taking all these tests together, it is clear that equations (2)–(4) imply the rejection of one aspect of the "cost-push" extreme view and to the nonrejection of the "monetarist" extreme view.

On the determination of excess demand, two things need to be established, the properties of the world demand for money function and the determinants of the world money stock. If there exists a stable world aggregate demand function for real balances, then one further aspect of the "monetarist" extreme position cannot be rejected. If, further, it can be shown that the direction of causation runs from money to prices and not vice versa, then the "cost-push" view has to be rejected as failing every test which may confront it while the "monetarist" position stands, pending the specification and performance of yet more searching tests.

Gray et al. (1976) estimated demand for money functions for G-10 of the following form:

(5) $\quad (m^* - p) = k + \beta y + \gamma r$

(6) $\quad \Delta(m - p) = \theta[(m^* - p) - (m_{-1} - p_{-1})] + u$

where m is the natural logarithm of the nominal money stock, r is the rate of interest, an * denotes desired, and the following sign restrictions apply to the parameters; $\beta > 0$, $0 \geq \theta \geq 1$, $\gamma \leq 0$. Gray et al. report the following estimates of the parameters in (5)–(6) on quarterly G-10 data[6] for 1957–71 (ratios of asymptotic standard errors to parameters in parentheses)

$k = 2.20 \quad \beta = 0.53 \quad \theta = 0.42$
$[13.92] \quad [25.24] \quad [5.45]$

with $R^2 = 0.986$.

These parameters were estimated by imposing $\gamma = 0$ and a parameter of first order autocorrelation $\rho = 0$. Freely estimating these parameters changes the estimated values of k, β and θ only slightly and the hypothesis that γ and ρ take on the extreme values imposed cannot be rejected at the 95 percent level. Further, a permanent income formulation was rejected in favor of that reported above. However, it is worth noting that permanent income and partial adjustment can only be discriminated between on the basis of error structures and so the above results might alternately be interpreted as representing a permanent income demand for money function. Thus, the hypothesis that there exists a stable world demand function for real balances which is interest inelastic cannot be rejected by the data for this period.[7] These results imply that one further aspect of the monetarist position cannot be rejected and that the world aggregate effect of national fiscal policies is primarily on real interest rates and not on output and prices.

The last remaining matter which needs attention before the "cost-push" extreme can be completely disposed of concerns the direction of causation. It remains a possibility that inflation is caused by some (exogenous to the model) cost-push factors with money passively responding to inflation. If that is the case, equa-

[6] For full details of data sources and methods, and estimation procedures, see Gray *et al.* The m variable is narrow, $M1$, money, income is real GNP aggregated over G-10 with quarterly data based on linear interpolation of annual data and the interest rate is that on Eurodollars.

[7] Splitting the data period between the (overlapping) first and last 40 observations reveals considerable structural stability although, for the last 40 observations, the interest rate does become significantly nonzero; its coefficient estimate is $-.034$ with an asymptotic standard error of 0.016.

tion (2) or (3) has to be interpreted as determining excess demand in a manner analogous to that suggested by Irving Fisher, and, given that level of demand and exogenously given rate of inflation, the demand for money function determines the money stock. This reverse causation story immediately runs into a difficulty given the time lags involved in (2) and (3). It is excess demand lagged one quarter which is best correlated with the difference between actual and expected inflation. Thus, for the reverse causation story to make sense it would have to be argued that excess demand during the current quarter is caused by a discrepancy between acutal and expected inflation which is not revealed until the next quarter. The timing makes life difficult for those who embrace the Fisher aggregate supply interpretation of the relation between excess demand and unanticipated inflation. It is also probably enough to discredit the reverse causation cost-push view If it is not, there is yet a further body of evidence which goes against it. Hans Genberg and Alexander Swoboda (1975), using the test suggested by Christopher A Sims show that "changes in the world money stock have, on the average preceded changes in both world income and the world price level during the last decade and a half."[8]

If the above results are brought together, they constitute a simple, yet complete model of the determination of the rate of inflation and the level of real output. The basic structure, repeated here for convenience is:

$$(7)\qquad \Delta p = \alpha x_{-1} + \Delta p^e$$

$$(8)\qquad \Delta p^e = \lambda \Delta p_{-1} + (1-\lambda)\,\Delta p^e_{-1}$$

$$(9)\qquad \Delta(m-p) = \theta[(k+\beta y) - (m_{-1} - p_{-1})$$

$$(10)\qquad y = y^* + x$$

where all the variables are as already defined and where y^* is the natural logarithm of "full employment" real output.[9] Solving these equations for Δp and x gives the following:[10]

$$(11)\quad \Delta p = \left(2 - \frac{\alpha}{\theta\beta}\right)\Delta p_{-1} - \left[1 - \frac{\alpha}{\theta\beta}(2-\lambda-\theta)\right]\Delta p_{-2} - \frac{\alpha}{\theta\beta}(1-\lambda)(1-\theta)\,\Delta p_{-3} + \frac{\alpha}{\theta\beta}\,\Delta m_{-1} - \frac{\alpha}{\theta\beta}(2-\lambda-\theta)\,\Delta m_{-2} + \frac{\alpha}{\theta\beta}(1-\lambda)(1-\theta)\,\Delta m_{-3}$$

$$(12)\quad x = \left(2 - \frac{\alpha}{\theta\beta}\right)x_{-1} - \left[1 - \frac{\alpha}{\theta\beta}(2-\lambda-\theta)\right]x_{-2} - \frac{\alpha}{\theta\beta}(1-\lambda)(1-\theta)x_{-3} + \frac{1}{\theta\beta}\,\Delta m - \left(\frac{2-\theta}{\theta\beta}\right)\Delta m_{-1} + \left(\frac{1-\theta}{\theta\beta}\right)\Delta m_{-2}$$

Given the particular values of α, β, λ, θ reported above, these two third-order difference equations generate a stable cyclical approach to the steady states (ignoring Δy^*) of $x^* = 0$ and $\Delta p^* = \Delta m$ with heavily damped cycles the period of which is twenty-four quarters. However, for a variety of reasons,[11] the equations are not suitable for a direct simulation test and probably would not track the history of p and x in a very close manner if dynamic simulation were performed Nevertheless, the structural equations estimated do make it impossible to reject the basic "monetarist" explanation of world average inflation and do provide a set of simple reduced-form equations for prices and output which qualitatively have properties

[8]Genberg and Swoboda (1975, p 21).

[9]Models with similar structures and properties to this have been suggested by Laidler (1973) and John Vanderkamp.

[10]Equations (11) and (12) are different from those reported in Parkin (1976) The equations here are correct and those in my earlier paper contained an error in the specification of the demand for money function which lead to an error in the difference equations

[11]The key reasons for this are that y in the reported demand for money function was based on annual national income accounts with quarterly data obtained by linear interpolation while x in the price equation was based on deviations from trend in a quarterly G-10 industrial production index. These inconsistencies are being adjusted in work currently underway

which the world clearly displays. Full employment ($x = 0$) and proportionality of inflation to money supply growth ($\Delta p = \Delta m$) are only steady state properties of the model advanced and any change in the rate of monetary expansion will be accompanied first by a change in the level of real economic activity and subsequently by a change in the inflation rate.

The discussion so far has treated the money supply as given and not enquired into the process whereby it is generated. Addressing this matter, Parkin et al. (1975) developed a simple model of the world money supply and, based on quarterly G-10 data from 1961 to 1971 concluded that there existed a stable, interest inelastic relation between the world money stock and the world monetary base, the latter defined as the sum of national monetary bases, with a secular increase in the broad money multiplier which they attributed to a gradual adjustment process in working off excess reserves. They also suggested that, up to 1968, the growth in domestic credit had been the main source of base growth but that after 1968, the growth of international liquidity began to dominate. However, they suggested that the relation between international reserves and total world base money was weak and therefore not exploitable for purposes of world monetary control.

More recent work makes it necessary to reevaluate and modify some of these conclusions. First, Swoboda (1975a) and Genberg and Swoboda (1976) show that there is an asymmetry in the effects of United States base and "rest of world" base on the world money supply arising from the fraction of reserves which the rest of the world holds as deposits with *U.S.* commercial banks and *U.S.* treasury bills as opposed to deposits with the Federal Reserve Banks. This makes *U.S.* base a "super-high-powered" money and predicts a larger multiplier effect of *U.S.* base on the world money stock than the bases of other countries. Empirical work performed by these authors confirms their proposition. Secondly, the connection between international liquidity and the world money supply has recently been thoroughly investigated by H. Robert Heller. He shows that there is a well determined distributed lag relation between these two variables with changes in liquidity clearly preceding changes in the world money supply.

The tentative conclusion which emerges from these studies is that the growth in the world money supply has been dominated by, though not completely determined by, the growth of international liquidity and that the growth of *U.S.* base money has been a key contributor to that liquidity and world money supply growth.

II. National Price Levels and the International Transmission Process

There are two basic hypotheses concerning the international transmission of inflation under fixed exchange rates both of which have been attributed to Hume. One is that the law of one price has no respect for national boundaries, hence, any discrepancies between prices of similar goods in different countries will quickly be arbitraged away ensuring equality of prices and of rates of price change across countries. It is recognized that measured rates of inflation do differ but suggested that such differences are attributable to trend changes in relative prices arising from different underlying productivity growth rates. Given a national rate of inflation arising from arbitrage, payments balances will ensure that the exogenous world money supply is distributed (endogenous) to each country to validate its price level behavior. There is an important variant of this hypothesis of the international transmission mechanism which separates traded from nontraded goods and has international arbitrage equalizing traded goods prices only with competitive labor and domestic goods markets bringing the inflation rates of wages and nontraded goods prices into alignment with (but not in general to equality with) the world rate of inflation. (See especially G. Edgren, K. O. Flaxen and G. E. Ohdner, and Parkin 1972, 1974.)

The second mechanism is one which has a rise in the world inflation rate leading to a fall in the relative domestic price level which generates excess demand and a balance of payments surplus. The excess demand and the money supply growth induced by the balance of payments surplus generate a process of rising domestic prices which continues until equilibrium has been restored. A variety of alternative *domestic* transmission mechanisms are compat-

ible with this international transmission mechanism including those of the Keynesian-Phillips curve and the mechanical quantity theory (See William H. Branson.)

The two international transmission mechanisms of inflation are clearly not mutually exclusive and a more general formulation would combine them. The simplest way of doing this is to adapt the standard expectations-augmented excess-demand model of price determination to the open fixed exchange rate economy. This has been done in a series of papers by Rodney Cross and Laidler, Laidler (1976), Franco Spinelli, Andrew Horseman, and Parkin et al. (1976). These studies postulate that domestic prices respond to domestic demand and to the expected rate of inflation where the latter variable is dependent not only on the history of domestic but also of world inflation If the pure arbitrage story is a good approximation to the world then domestic excess demand will be unimportant in determining domestic price change, and world inflation will be the only influence on inflation expectations. If the other transmission mechanism is the only relevant one then domestic excess demand and inflation expectations based only on domestic considerations will dominate and world inflation will have no important separate influence on the domestic inflation rate.

The broad consensus of the empirical work performed and reported in the studies cited above is that neither extreme is an adequate simplification and that there are elements of both in the generation of national price level movements, at least as far as quarterly and annual averages are concerned Over longer time averages, the arbitrage process seems much stronger (see Genberg, 1976) and no studies are available which suggest that foreign prices can be ignored when explaining price movements in open economies.

Those studies which have permitted expectations of world inflation as well as of domestic inflation to have a direct effect on domestic inflation have an important implication for inferences concerning the existence or otherwise of a long-run tradeoff between inflation and unemployment at the national level. The dominant finding of the studies just cited is that no long-run tradeoff exists. Parkin and Graham Smith show that earlier studies which did display a long-run tradeoff can be reconciled with those that do not by analyzing the consequences for parameter estimates of the omitted world inflation variable.

A further matter on which these studies shed some light concerns the specific foreign prices which have the main direct impact on domestic prices. Three broad possibilities have been suggested. Early postwar studies emphasized the role of import prices; the "Scandinavian" approach emphasizes export prices while the arbitrage approach suggests a broad index of all foreign prices. Studies can be found which show all to be important and the only direct attempt to compare some of the alternatives, by Laidler (1976), suggests that a broader index is better than a narrower one.

All the studies referred to above deal only with the proximate determinants of prices. Laidler (1975) has incorporated a price setting mechanism of the above type into a complete macro model and shown that, while there is no ambiguity that a rise in the world inflation rate raises domestic inflation, there *is* an ambiguity about its impact effect on domestic output which could fall if the impact on the price level exceeds that on the money stock. Spinelli has applied the Laidler model to the Italian economy and found it to have a high degree of explanatory power and to completely outperform an alternative "cost-push" explanation for that country. The alternative "cost-push" explanation of inflation at the level of the individual country has futher been investigated by Peter D. Jonson for Australia and George Zis for the Group of Ten and again shown to be easily rejected.

III. Concluding Remarks

The preceding summary account of theoretical and empirical work on the generation and transmission of inflation in a fixed exchange rate world does not itself require a summary. It is worthwhile, however, to try to highlight the outstanding issues which need further attention. First, a fully consistent structural model of world average inflation capable of providing a close dynamic tracking of world output and price level movements remains to be built. Se-

condly, the precise effects on the world money supply of domestic credit creation in the United States and in the smaller countries need to be further clarified. Thirdly, the details of the international and domestic transmission mechanisms need further specification. A key to the advancement of knowledge in these last areas will be the explicit *comparison* of competing hypotheses, all too little of which has been undertaken to date.

REFERENCES

Jean Bodin, "Response aux paradoxes de Malestroit touchant l'Encherissement de toutes Choses et le Moyen d'y remedier," English trans is Arthur E. Monroe, ed., *Early Economic Thought*, Cambridge, Mass. 1924.

William H. Branson, "Monetarist and Keynesian Models of the Transmission of Inflation," *Amer Econ Rev. Proc.*, May 1975, *65*, 115–19.

Emil Claassen and Pascal Salin, *Recent Issues in International Monetary Economics*, Amsterdam 1976.

Rodney Cross and David Laidler, "Inflation, Excess Demand and Expectations in Fixed Exchange Rate Open Economies: Some Preliminary Empirical Results," in M Parkin and G. Zis, eds , *Inflation in the World Economy*, 1976a, 221–54.

Nigel Duck, Michael Parkin, David Rose and George Zis, "The Determination of the Rate of Change of Wages and Prices in the Fixed Exchange Rate World Economy, 1956–1971," in Parkin and Zis, 1976a, 113–42.

G. Edgren, K. O. Flaxen and G. E. Ohdner, "Wages Growth and the Distribution of Income," *Swedish J. Econ.*, 1969, *71*, 133–60

Irving Fisher, *The Purchasing Power of Money*, New York 1911 and 1963.

Jacob A. Frenkel, "Adjustment Mechanisms and the Monetary Approach to the Balance of Payments," in E. Claassen and P. Salin, 1976.

Hans Genberg and Alexander K. Swoboda, "Causes and Origins of the Current Worldwide Inflation," Discussion Paper, Ford Foundation International Monetary Research Project, Graduate Institute of International Studies, Geneva, Nov. 1975.

——— **and** ———, "Worldwide Inflation Under the Dollar Standard," mimeo, Graduate Institute of International Studies, Geneva 1976.

———, "A Note on Inflation Rates Under Fixed Exchange Rates," in Parkin and Zis, 1976a, 183–88

Malcolm R. Gray and Michael Parkin, "Discrimination Between Alternative Explantions of Inflation" in Michele Fratianni and Karel Tavernier, eds , Proceedings of Conference on "Money, Banking and Credit in Open Ecomomies, Catholic University of Louvain, 1973.

———, **R. Ward and George Zis,** "The World Demand for Money Function: Some Preliminary Results," in Parkin and Zis, 1976a, 151–77.

H. Robert Heller, "International Reserves and World Wide Inflation," *I.M.F Staff Paper*, Mar. 1976, *23*, 61–87.

Andrew Horseman, "The Relation between Wage Inflation and Unemployment in an Open Economy," in M. Parkin and G Zis, *Inflation in Open Economies*, 1976b, 175–200

David Hume, "Of Money," and "Of the Balance of Trade," in *Essays, Moral, Political and Literary*, 1741, Oxford 1963, 289–302, 316–33.

Harry G. Johnson, "Inflation and the Monetarist Controversy," *Professor Dr. F. de Vries Lectures*, Amsterdam 1972.

———, "World Inflation and the International Monetary System," *The Three Banks Review*, Sept. 1975, no. 107, 3–22.

Peter D. Jonson, "World Influences on the Australian Rate of Inflation," in Parkin and Zis, 1976b, 237–58.

A. B. Laffer and D. I. Meiselman, *The Phenomenon of Worldwide Inflation*, American Enterprise Institite for Public Policy Research, Washington 1975.

David Laidler, "The Influence of Money on

Real Income and Inflation: A Simple Model with Some Empirical Tests for the United States, 1953–1972," *Manchester School*, Dec. 1973, *41*, 367–95.

———, *Essays on Money and Inflation*, Ch. 9, "Price and Output Fluctuations in an Open Economy," Manchester 1975, 183–94.

———, "Alternative Explanations and Policies Towards Inflation: Tests on Data Drawn from Six Countries," forthcoming in K. Brunner and A. H. Meltzer, eds., Carnegie-Rochester Conference Series on Public Policy, 1976.

——— **and Michael Parkin,** "Inflation: A Survey," *Econ J.*, Dec. 1975, *85*, 741–809.

——— **and A. R. Nobay,** "Some Current Issues Concerning the International Aspects of Inflation" in Claassen and Salin, 1976.

R. Mundell, *Monetary Theory: Inflation, Interest and Growth in the World Economy*, Goodyear 1971.

Michael Parkin, "Inflation, The Balance of Payments, Domestic Credit Expansion and Exchange Rate Adjustments," in Robert Z. Aliber, ed , Proceedings of the Conference on National Monetary Policies and the International Financial System, Racine 1972.

———, "World Inflation, International Relative Prices and Monetary Equilibrium Under Fixed Exchange Rates," paper presented at the Conference on the Political Economy of Monetary Reform, Racine 1974, and forthcoming in the Proceedings of that conference.

———, "International Liquidity and World Inflation in the 1960s," in Parkin and Zis, 1976b, 48–63.

——— **Ian Richards and George Zis,** "The Determination and Control of the World Money Supply Under Fixed Exchange Rates, 1961–1971," *The Manchester School*, Sept. 1975, *43*, 293–316; reprinted in Parkin and Zis, 1976b, 24–47.

——— **and Graham W. Smith,** "Inflationary Expectations and the Long-run Trade-off Between Inflation and Unemployment in Open Economies," in Parkin and Zis, 1976b, 280–90.

———, **M. T. Sumner and R. Ward,** "The Effects of Excess Demand, Generalized Expectations and Wage-Price Controls on Wage Inflation in the U.K.," in K. Brunner and A. H. Meltzer, *Proceedings of the Conference on Wage-Price Controls*, Rochester 1973, Amsterdam 1966.

——— **and George Zis,** eds., *Inflation in the World Economy*, Manchester and Toronto 1976b.

——— **and** ———, eds., *Inflation in Open Economies*, Manchester and Toronto, 1976b.

Edward S. Shaw, "International Money and International Inflation: 1958–1973," *Business Review*, Spring 1975, Federal Reserve Bank of San Francisco, 5–17.

Christopher S. Sims, "Money Finance and Causality," *Amer. Econ. Rev.*, 1972, *62*, 540–52.

Franco Spinelli, "The Determinants of Price and Wage Inflation: The Case of Italy," in Parkin and Zis, 1976b, 201–36.

Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976.

Alexander K. Swoboda, "Gold, Dollars, Euro-dollars and the World Money Stock," Discussion Paper, Ford Foundation International Monetary Research Project, Graduate Institute of International Studies, Geneva, Nov. 1975a.

———, "Monetary Approaches to the Transmission and Generation of Worldwide Inflation," Discussion Paper, Ford Foundation International Monetary Research Project, Graduate Insitute of International Studies, Geneva, Mar. 1975b.

R. L. Teigen, "Interpreting Recent World Inflation," *Amer. Econ Rev. Proc.*, May 1975, *65*, 129–32.

John Vanderkamp, "Inflation: A Simple Freidman Theory with a Phillips Twist," *J. Monetary Econ.*, Mar. 1975, *1*, 117–22.

H. Johannes Witteveen, "Inflation and the International Monetary Situation," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 108–*14*.

George Zis, "Inflation: An International Monetary Problem or a National Social Phenomenon," in Parkin and Zis, 1976b, 1–23.

MONETARY THEORY FOR OPEN ECONOMIES: THE STATE OF THE ART

Micro Theory of International Financial Intermediation

By CHARLES FREEDMAN*

In this paper I examine the microeconomic theory of financial intermediation in an open economy. Because of the paucity of literature on international financial intermediation, much of the emphasis is necessarily placed on models of domestic intermediation and the possibility of extending them to the international economy.

I. A Taxonomic Approach

International financial intermediation can be defined to include three main types of transactions.

1) Financial transactions of a bank (a term used henceforward to include all intermediaries since banks are by far the major participants in these activities) with nonresidents denominated in the currency of the country in which the bank is resident. This item includes, for example, *U.S.* dollar deposits at American banks by nonresidents of the United States and *U.S.* dollar loans by American banks to nonresidents of the United States.

2) Financial transactions of a bank with nonresidents denominated in currencies other than the currency of the country in which the bank is resident. This category includes, for example, Eurocurrency deposits and Eurocurrency loans made by Eurobanks with nonresidents of the country in which the Eurobank is resident.[1]

3) Financial transactions of a bank with residents denominated in foreign currencies. This category includes, for example, *U.S.* dollar deposits of Canadian residents at Canadian banks and *U.S.* dollar loans to Canadian residents by Canadian banks.

Making use of these distinctions, one can construct a balance sheet of an individual bank or some aggregate of banks involved in international financial intermediation as follows

Assets	*Liabilities*
Claims on residents, domestic currency	Liabilities to residents, domestic currency
Claims on foreigners, domestic currency	Liabilities to foreigners, domestic currency
Claims on residents, foreign currency	Liabilities to residents, foreign currency
Claims on foreigners, foreign currency	Liabilities to foreigners, foreign currency

According to the definition introduced above,

*Research Adviser, Research Department, Bank of Canada. The views expressed in this paper are those of the author and no responsibility for them should be attributed to the Bank of Canada

[1]One might wish to distinguish further by the country of ownership of the Eurobank Thus Eurodollar loans by London branches of American banks might be treated differently for some purposes from Eurodollar loans by London branches of German banks The distinction is important if it is believed that the head office of the bank makes decisions based on the consolidated balance sheet of the entire bank Similarly, one might wish to distinguish between Eurodollar deposits of American residents and those of German residents, etc

the term international financial intermediation includes the bottom three items on each side of the balance sheet. In addition, one can define the net foreign currency asset position as the difference between the bank's claims in foreign currency (on both residents and nonresidents) and the bank's liabilities in foreign currency (to both residents and nonresidents). This is a measure of the bank's net spot position in foreign currency and movements in this measure ought to be explained by a theory of financial intermediation.[2]

There are at least four levels from which one can approach the theory of financial intermediation. First, there is the level of all the intermediaries involved in a particular market. Thus, for example, in the case of the Eurodollar market, this would include all banks involved in collecting Eurodollar deposits and making Eurodollar loans. At this level of analysis the most useful approach is the macroeconomic theory of financial intermediation, derived from the work of James Tobin and William Brainard and Tobin (1969). The application of this framework to the Eurodollar market is carried out in John Hewson and Eisuke Sakakibara and Charles Freedman (1977a). This type of model yields insights into the effects of autonomous shifts and policy changes on the Eurodollar interest rate, the balance of payments, domestic interest rates, and the size of the Eurodollar market. The effect of various kinds of controls with which governments have attempted to modify the influence of the Eurodollar market on their economies can also best be examined with the macroeconomic model of financial intermediation.

The next two levels of analysis are based on the nationality of the banks. Here there are two alternatives. One can focus either on all the banks engaged in international financial intermediation which are resident in a given country, regardless of the location of head office, or one can focus on all the banks controlled from a given country, regardless of the residence of the branches. There are a number of reasons for analyzing the behavior of banks at one or the other of these national levels. First, controls are normally imposed on banks resident within a given jurisdiction (e.g., controls on net foreign currency asset position) or affect the behavior of banks controlled from a given jurisdiction (e.g., reserve ratios on U.S. bank borrowing from the Eurocurrency market which influenced the behavior of the subsidiaries of U.S. banks in London). Second, political risk considerations can affect the interest rate on Eurodollar deposits at banks in a given country (Robert Aliber 1973) and default risk considerations can affect the rate on deposits at banks controlled in a given country. Third, the data are sometimes available by country of residence as in the case of the Eurobanks operating in the United Kingdom and sometimes by country of control as in the case of the Canadian banks. The classification by country of control is more useful to the extent that head office controls all its subsidiaries and acts on its total portfolio regardless of where booked. It is less useful to the extent that subsidiary banks operate relatively autonomously and have control of their own balance sheets. Analysis at the national level can be either macroeconomic or microeconomic in spirit. Examples of the latter are the studies by Freedman (1974) on the Canadian banks and Ralph Bryant and Patric Hendershott on borrowing from the United States by Japanese banks.

The fourth level of analysis is microeconomic and focuses on the behavior of the individual banks and that of the other participants in financial transactions. The first set of questions in the microeconomic approach relates to the gains from financial intermediation to all three participants in the act of financial intermediation—the ultimate lender, the ultimate bor-

[2] The net foreign currency asset position of the banks arises from conversions by the banks between Eurocurrency liabilities and domestic currency assets or domestic currency liabilities and Eurocurrency assets. It is to be distinguished from the net claims of the banks on nonresidents (in both domestic and foreign currencies). The latter item is not of interest to the banks unless it is the object of government regulation. See Rodney Mills.

rower, and the financial intermediary. A second aspect of this type of analysis lies in the modeling of the behavior of the intermediary. In some models the focus is on the transactions costs of intermediation, in others on the rising demand curve for deposits at the intermediary and the interest rate setting behavior of intermediaries, in yet others the risk factors predominate. It is the microeconomic level of analysis that I will now examine in detail.

II. The Gains from Intermediation

In the domestic economy, the question as to why intermediaries exist can be put as follows: What advantages are there to the ultimate lender and ultimate borrower from intermediated finance as opposed to direct finance and whence derives the profit margin of the intermediary? The answers to these questions were originally given by John Gurley and Edward Shaw. The advantage to both the lender and borrower is that the instruments created by the intermediary (e.g., deposits, mortgage loan) are better suited to their requirements in terms of liquidity, maturity, etc. than the instrument used in direct finance. In return for a more suitable instrument, the depositor is willing to accept a lower rate of return and/or the borrower is willing to pay a higher rate of interest than would be the case in direct finance.[3] This opens up a profit margin for the intermediary sufficient to induce him to take on the risks of intermediation.[4] At a slightly deeper level of analysis, one may inquire as to why the intermediary is able to do what the ultimate lender and borrower cannot do for themselves. The answer is that economies of scale in lending and borrowing permit the reduction of risks and costs of several kinds. 1) Diversification on the asset side of the intermediary's balance sheet reduces default risk. 2) The statistical predictability of the redemption of intermediary liabilities enables the intermediary to engage in maturity transformation, i.e., the intermediary assets are less liquid than its liabilities. 3) Specialization in extending loans reduces costs and risks of default. 4) Transactions costs (including costs of information) in bringing together lenders and borrowers are reduced.[5]

Now in extending this type of analysis to international financial intermediation, one has to put in a slightly different form the question of why financial intermediaries exist. What advantages are there to ultimate lenders and borrowers from international financial intermediation as opposed to domestic direct finance, international direct finance, and domestic financial intermediation? Whence derives the profit to the intermediary in international financial intermediation? In a comparison of international financial intermediation to domestic and international direct finance, the same considerations should apply as in the comparison of domestic financial intermediation and domestic direct finance. That is, there are various advantages derived from economies of scale in lending and borrowing, namely diversification, reduction of withdrawal risks, specialization, and reduction of transactions costs.[6]

The comparison of international financial intermediation with domestic financial intermediation is of greater analytical interest. To induce lenders to shift from claims on domestic financial intermediaries to claims on banks denominated in foreign currencies or claims on banks in foreign countries it is generally necessary to offer some motivation to offset the added risks.[7]

[3]If the transactions costs of direct finance are sufficiently high it is possible for the net return to the lender to be higher, and the net cost to the borrower lower, in the case of intermediated finance than in the case of direct finance.

[4]For a formal treatment of this type of analysis, see I. D. Mangoletsis.

[5]For detailed discussions of these points see Gurley and Shaw, or Charles Goodhart.

[6]It might be argued that the difficulties in obtaining information on the credit standing of potential borrowers in international lending would be a factor in the development of international financial intermediation. However, since most borrowers in the Eurodollar market are prime names it is not likely that this is a major consideration.

[7]One exception to this is the case of foreign currency working balances held by international traders.

Such motivation usually takes the form of higher deposit rates or better terms (e.g., shorter maturities) than those offered at domestic financial intermediaries. Similarly, to attract ultimate borrowers to loans denominated in foreign currency or loans at foreign banks, it is generally necessary to offer lower loan rates or better terms than those offered domestically. Thus gross profit margins in international financial intermediation must be less than those in domestic financial intermediation although sufficiently large to induce the intermediaries to expand their activities in this sphere.

These considerations give rise to the following interpretation of the rapid growth in recent years of international financial intermediation vis-à-vis domestic financial intermediation. Because of legal restrictions (ceilings on deposit rates, reserve requirements) or conventional cartel type arrangements (in a banking system with barriers to entry), domestic financial intermediaries have maintained gross profit margins greater than those needed to cover operating costs and earn a normal profit. The competitive response to these large profit margins has been the development of international financial intermediation which has attracted deposits and loans away from domestic intermediation by offering rates based on lower profit margins.[8]

If this interpretation is correct one can then ask how domestic financial intermediaries have continued to coexist with international financial intermediaries when the latter have offered higher deposit rates and lower loan rates than the former. To this question there are at least three answers. 1) There are certain types of liabilities (e.g., checkable deposits used as the means of payment) and certain types of assets (e.g., mortgage loans) that are not offered by international financial intermediaries. 2) To maintain the profitability of their business in spite of low gross profit margins, international financial intermediaries must keep operating costs low. Hence the market handles only wholesale transactions in deposits of very substantial size (typically $1 million according to Bell) and unsecured loans to prime name borrowers.[9] 3) Because of the political risk in holding deposits at foreign banks (Aliber 1973) considerations of risk reduction entail that at least some deposits be held in the risk-free domestic financial intermediary.[10]

III. Models of Intermediary Behavior

In recent years there have been a number of studies that have modeled the financial intermediary as a firm maximizing some form of objective function. For our purposes these studies can be divided into five groups: 1) studies of a bank whose deposits (whether fixed or stochastic) are beyond its control; 2) studies of a savings and loan association whose assets have much longer maturity than its liabilities; 3) studies of a financial intermediary which has control over its deposit rate and/or its loan rate; 4) studies of a financial intermediary facing given deposit and loan rates whose growth is limited by the increasing marginal cost of attracting deposits and/or loans; 5) studies of a financial intermediary facing given deposit and loan rates whose growth is limited by risk considerations. Only the last three approaches have been useful in explaining the behavior of international financial intermediaries.

[8] For a more detailed analysis of the ability of Eurobanks to operate on smaller margins than domestic banks see Geoffrey Bell. Alexander Swoboda offers an alternative interpretation of the development of the Eurodollar market in terms of the reduction of transactions costs connected with the growing use of "vehicle currencies." Jurg Niehans and Hewson explain the size of the Eurodollar market in terms of the different transactions costs of different kinds of banks as expressed in the difference between borrowing and lending rates for different customers. For a detailed discussion of the growth of the Eurodollar market which emphasizes government controls and other shocks, see Freedman (1977b).

[9] If the international financial intermediaries are skimming off the large deposits and loans from domestic banking systems, then the average size of deposit and the average size of loan in the latter should be growing more slowly than, say, wealth per capita or average firm size. This is a testable hypothesis that can perhaps be investigated with national data.

[10] Alternatively, deposits could be spread among banks in several countries.

1) An excellent recent treatment of the microeconomic behavior of banks facing externally determined deposits is to be found in Donald Hester and James Pierce.[11] They carefully set out a number of distinct models of portfolio behavior, focusing on many aspects of banking that might be expected to affect portfolio composition. Although some of these aspects might be relevant to a study of the behavior of an international financial intermediary, the key assumption made throughout by Hester and Pierce, namely that the individual bank has virtually no control over its deposits, makes the models unsuitable for extension to the international setting.

2) In a series of articles by Gerald Weber, Paul Meyer, Stephen Goldfeld and Dwight Jaffee, Patric Hendershott, and George Daly, the deposit-rate setting behavior of a savings and loan association has been analyzed in detail. In general, these articles assume that the intermediary has control over the interest rate paid on its deposits and faces a rising demand curve for deposits. The crucial point of the institutional setting is that in the United States, mortgages (the asset in which saving and loan associations invest the bulk of their funds) have a much longer maturity than do savings and loan association deposits. These models also assume that there is no secondary market in mortgages. Since most loans made by Eurobanks have interest rates that are adjusted periodically to deposit rates and since, according to Hewson, the maturity structure (at least as defined in terms of the period over which the interest rate is adjusted as opposed to the loan commitment period) of Eurobanks is nearly balanced, this set of articles also is not suitable for extension to the international setting.

3) The third set of studies deals with a financial intermediary which has control over its deposit and/or loan rates and faces an upward sloping demand curve for deposits and/or a downward sloping supply curve of loans. The intermediary is assumed to maximize profits or size or an objective function based on both profits and size or the expected rate of return on equity. Both Mario Monti (1971, 1972) and Myron Slovin and Marie Sushka examine the differing implications of whether profits or size or both appear in the bank's objective function. The former focuses mainly on bank responses to changes in monetary policy of different types whereas the latter focuses mainly on the response of the intermediary's deposit rate to changes in competing market rates. Michael Klein and Richard Towey also introduce service charges and the cost of administering the payments mechanism into their models.

The extension of this type of model to the international economy has been carried out in Freedman (1974). The Canadian banks (treated as a single bank) are assumed to face an upward sloping demand curve for foreign currency deposits[12] from each of three types of depositors (Canadian, American, rest of world). The interest rate on deposits is set so as to maximize profits given the interest rate on foreign currency assets in which the banks invest. That is, the marginal cost of an additional dollar of foreign currency deposits is equated to the marginal revenue from an additional dollar of foreign currency assets. The interest rate on deposits, together with the actual demand curves for deposits, determines the amount of deposits forthcoming.

4) In the fourth type of study, John Kareken focuses on a bank which takes interest rates on both deposits and loans as given. That is, there is an infinitely elastic demand for deposits at the bank and an infinitely elastic supply of loans to the bank. However, the bank has a rising marginal cost associated with increased deposits and a falling marginal revenue associated

[11] There is a useful survey of earlier studies of bank behavior when deposits are externally determined in Chapter 1 of this book.

[12] The finiteness of the elasticity of demand derives from the fact that depositors perceive foreign currency deposits at Canadian banks as imperfect substitutes for foreign currency deposits at other banks. This perception derives mainly from the differences in political risk and default risk among different banks.

with increased loans. Either one of these conditions is sufficient to yield a determinate equilibrium for the bank where marginal revenue is equal to marginal cost.

Hewson, Chapter 4, extends this type of model to the international economy. The Eurobank is assumed to face given interest rates on loans and deposits. It is further assumed that there are increasing marginal transactions costs on both loans and deposits[13] and that the bank maximizes profit net of transactions costs. In equilibrium deposits and loans are such that the interest rate differential between the rates on deposits and loans exactly covers the marginal transactions costs.

5) The fifth type of model focuses on the increasing risks from intermediary expansion as the main reason for the limit on intermediary growth. Michael Parkin treats the *U.K.* discount houses (an intermediary operating in the domestic economy) as maximizing the expected value of a constant absolute risk aversion utility of profits function. This is shown to be equivalent to maximizing expected profits minus a constant times the variance of profits. With uncertainty about both asset yields and borrowing costs, expected profits are linear in the size of the portfolio vector and the variance of profits is quadratic in the size of the portfolio vector. Hence the limit on the size of the intermediary derives from the increase in the variance of profits relative to expected profits as the intermediary expands. Parkin also assumes the exogeneity of one item in the balance sheet.

A similar approach is used in Bryant and Hendershott to explain borrowing from the United States by Japanese banks. Using the Tobin-Markowitz framework the authors make all intermediary assets and liabilities a function of the interest rates on assets and liabilities, risk variables, and the scale variable, net worth. With the latter exogenous, expansion of both assets and liabilities increases risk and hence the size of the intermediary at any point of time is limited by the volume of net worth.[14]

Thus there are three main ways in the literature of determining the scale of an individual financial intermediary operating in the international economy and of explaining its behavior—noncompetitive setting of deposit and/or loan rate based on rising demand curve for deposits and/or falling supply curve of loans; rising marginal transaction costs on the part of the financial intermediary in extending deposits and/or loans; increasing risk as the intermediary expands, perhaps resulting from the exogeneity of one item in the balance sheet. The choice among these approaches or the use of some model incorporating more than one of them must be made on empirical grounds, not on theoretical grounds. In choosing among them it would help if both the assumptions and implications of each approach were analyzed more carefully. How elastic are the demand curves for deposits? What evidence is there for rising marginal cost curves? Why should one item on a balance sheet be exogenous? One important implication of each model is its explanation of the growth of the intermediary. In the first approach, the scale variables of the ultimate lender and borrower are assumed to grow. In the second approach, transaction costs fall or the differential between deposit and loan rates increases or growth comes from the establishment of new intermediaries. In the third approach, it is the growth of the exogenous item that leads to growth of the intermediary.[15]

One final aspect of the microeconomic ap-

[13] In the more complicated version of the model, there are also costs associated with the maturity mismatch of assets and liabilities.

[14] For the sake of completeness one should mention the articles by David Pyle and Oliver Hart and Dwight Jaffee which apply portfolio selection theory to financial intermediaries. Pyle examines the circumstances under which intermediation is likely to arise. Hart and Jaffee prove the existence of a separation theorem for financial intermediaries and then derive the comparative static properties of their model. There has been no extension of these models to the international economy.

[15] In Bryant and Hendershott it is the net worth item that is exogenous to the model. If expansion of the intermediary were profitable it would seem reasonable for the intermediary to issue more equity in order to permit such expansion.

proach to financial intermediation is the chain of bank deposits between the ultimate lender and the ultimate borrower. Aliber (1976) has argued that some intermediaries specialize in deposit gathering, others in making loans. The former will tend to redeposit their funds in other banks and the latter will tend to gather their funds from other banks.[16] Hewson, Chapter 6, has set up a small expository model in which such a result occurs because of different transactions costs for different banks on different types of transactions. Once again there is the need for further analytical and empirical work to explain why transactions costs differ among different types of transactions and different types of banks.

REFERENCES

Robert Z. Aliber, "The Interest Rate Parity Theorem: A Reinterpretation," *J. Polit. Econ.*, Nov.-Dec. 1973, *81*, 1451–59.

———, "Towards a Theory of International Banking," *Federal Reserve Bank of San Francisco Economic Review*, Spring 1976, 5–8.

Geoffrey Bell, *The Euro-dollar Market and the International Financial System*, New York 1973.

Ralph C. Bryant and Patric H. Hendershott, "Financial Capital Flows in the Balance of Payments of the United States: an Exploratory Empirical Study," *Princeton Studies in International Finance*, 25, Princeton 1970.

George G. Daly, "Financial Intermediation and the Theory of the Firm: An Analysis of Savings and Loan Association Behavior," *Southern Econ. J.*, Jan. 1971, *37*, 283–94.

Robert F. Emery, "The Asian Dollar Market," *International Finance Discussion Papers, 71*, Nov. 1975, Federal Reserve System.

Charles Freedman, "The Foreign Currency Business of the Canadian Banks: An Econometric Study," *Bank of Canada Staff Research Studies, 10*, 1974.

———, "A Model of the Eurodollar Market," *J. Monetary Econ.*, forthcoming, 1977a.

———, "Review Article on the Eurodollar Market," *J. Monetary Econ.*, forthcoming, 1977b.

Stephen M. Goldfeld and Dwight M. Jaffee, "The Determinants of Deposit-Rate Setting by Savings and Loan Associations," *J. Finance*, June 1970, *25*, 615–32.

Charles A. E. Goodhart, *Money, Information and Uncertainty*, London 1975.

John G. Gurley and Edward S. Shaw, *Money in a Theory of Finance*, Washington 1960.

Oliver D. Hart and Dwight M. Jaffee, "On the Application of Portfolio Theory to Depository Financial Intermediaries," *Rev. Econ. Stud.*, Jan. 1974, *41*, 129–47.

Patric H. Hendershott, "Financial Disintermediation in a Macroeconomic Framework," *J. Finance*, Sept. 1971, *26*, 843–56.

Donald D. Hester and James L. Pierce, *Bank Management and Portfolio Behavior*, New Haven 1975.

John Hewson, *Liquidity Creation and Distribution in the Eurocurrency Markets*, Lexington 1975.

——— **and Eisuke Sakakibara,** *The Eurocurrency Markets and their Implications*, Lexington 1975.

John H. Kareken, "Commercial Banks and the Supply of Money: A Market-Determined Demand Deposit Rate," *Fed. Reserve Bull.*, Oct. 1967, *53*, 1699–1712.

Michael A. Klein, "A Theory of the Banking Firm," *J. Money, Credit, Banking*, May 1971, *3*, 205–18.

I. D. Mangoletsis, "The Microeconomics of Indirect Finance," *J. Finance*, Sept. 1975, *30*, 1055–63.

[16]This is not unlike a domestic branch banking system in which some branches primarily gather deposits and others primarily make loans. The explanation for such behavior domestically lies in the difference in location of ultimate lenders and ultimate borrowers. An analysis of geographical aspects of behavior has been used by Robert Emery in his explanation of the growth of the Asian dollar market.

Paul A. Meyer, "Comment," *J. Finance*, Sept. 1966, *21*, 515–21.

Rodney H. Mills, Jr., "The Regulation of Short-Term Capital Movements in Major Industrial Countries," *Staff Economic Studies*, *74*, 1972, Federal Reserve System.

Mario Monti, "A Theoretical Model of Bank Behavior and Its Implications for Monetary Policy," *L'Industria Revista di Economia Politica*, April-June 1971, *2*, 165–91.

———, "Deposit, Credit and Interest Rate Determination under Alternative Bank Objective Functions," in G. P. Szego and K. Shell, eds., *Mathematical Methods in Investment and Finance*, Amsterdam 1972.

Jürg Niehans and John Hewson, "The Eurodollar Market and Monetary Theory," *J. Money, Credit, Banking*, Feb. 1976, *7*, 1–26.

Michael Parkin, "Discount House Portfolio and Debt Selection," *Rev. Econ. Stud.*, Oct. 1970, *37*, 469–97.

David H. Pyle, "On the Theory of Financial Intermediation," *J. Finance*, June 1971, *26*, 737–47.

Myron B. Slovin and Marie E. Sushka, *Interest Rates on Savings Deposits*, Lexington 1975.

Alexander K. Swoboda, "The Euro-Dollar Market: An Interpretation," *Essays in International Finance*, *64*, Princeton 1968.

James Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15–29.

——— **and William C. Brainard,** "Financial Intermediaries and the Effectiveness of Monetary Controls," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 383–400.

Richard E. Towey, "Money Creation and the Theory of the Banking Firm," *J. Finance*, Mar. 1974, *29*, 57–72.

Gerald S. Weber, "Interest Rates on Mortgages and Dividend Rates on Savings and Loan Shares," *J. Finance*, Sept. 1966, *21*, 515–21.

The Microeconomics of the Firm in an Open Economy

By MICHAEL ADLER AND BERNARD DUMAS*

This paper undertakes to survey a largely nonexistent literature. The extension of the theory of the firm to the open economy is still in its nascence. Consequently, this paper makes an attempt to fill in gaps by defining and discussing issues which have not yet received systematic treatment elsewhere.

The traditional theory of the firm assumed profit maximization which is Pareto-optimal under certainty. Early attempts to model the international firm (e.g., Guy V. G. Stevens, T. Horst and Adler and Stevens) all employed this paradigm. The neoclassical approach, however, breaks down under uncertainty. Profit maximization no longer ranks alternative decisions since profits are no longer under the firm's complete control. Value-maximization is then the most practicable option, despite the likelihood that it will not be unanimously preferred. The paper therefore examines the possibility for and nature of value maximizing decisions for open-economy firms.

The paper stops short of addressing empirical questions such as, say, the impact of a change in exchange risk on international corporate financial behavior and the implications for capital flows. Such issues are a matter of the comparative statics or dynamics of optimal decisions. Our concern is with the logically prior question of whether such decision rules exist and can be characterized.

I. The Objective of the Firm: General Considerations

A logical prerequisite to the computation of firms' decisions is the specification of an objective function which the firm can reasonably be postulated to maximize. It was long taken for granted that the corporate firm ought to maximize the sum of the market values of the securities it has issued and issues currently, and of the current cash flow distribution to its capital suppliers; an objective which we shall henceforth refer to as "value maximization." But since the firm is owned by stockholders, we must ask whether this objective is to their satisfaction. More precisely, since it is preferable not to have to resolve possible conflicts between stockholders, one might simply ask: When will the firms' decisions be Pareto-optimal from the standpoint of stockholders? Assuming that the welfare of nonstock securities holders is protected by covenants and the law,[1] we can rephrase the question: When will firms' decisions be Pareto-optimal from the standpoint of all investors?

After a long debate, the definitive answer to this question was provided by Niels Nielsen's dissertation.[2] According to this rendering, the main distinction to be introduced is whether or not the set of securities available in the capital market (for a given production decision of firms) affords an unconstrained Pareto optimal *(UPO)* allocation of consumption among households.

If the market is *UPO*, Nielsen proves two important theorems. First, the principle of Conservation of Investment Value *(CIV)* obtains; i.e., if one transforms linearly the cash flows of securities without affecting the total cash flows,

*Professor of Business, Columbia University and Assistant Professor, Columbia University (on leave) and E.S.S.E.C., France, respectively. Financial support from the Rockefeller Foundation is gratefully acknowledged.

[1]See Eugene Fama and Merton Miller, p. 151, and J. H. Scott.

[2]The interpretation of Nielsen's results provided here is solely our responsibility.

the market value of the new securities is equal to the same linear transformation of the old securities' market values. Second, value maximizing firms make Pareto optimal production decisions provided only that they behave as price takers with respect to the prices of existing securities. The *CIV* result provides us with sufficient conditions for cash flow preserving decisions to be irrelevant: among these are: the merger, debt-equity-mix and forward-contracting decisions as well as the currency-denomination *(c-d)* decision for the debt in the absence of taxes and bankruptcy costs. The second result provides us with a convenient maximand to reach the optimal level of these decisions which do modify total cash flows such as the production and investment decisions as well as the merger, debt-equity-mix, debt-denomination and forward-contracting decisions in the presence of taxes and bankruptcy cost.[3]

Unfortunately, the market is *UPO* under circumstances which cannot reasonably be expected to obtain. One sufficient condition for *UPO* is completeness in the Kenneth Arrow-Gerard Debreu sense but it would require a prohibitively larger number of securities and would imply astronomical market-making costs. Other sufficient condtions under which the market is *UPO* without being complete are spelled out in various so-called Portfolio Separation Theorems.[4] They usually assume that all investors share homogeneous probability beliefs regarding the purchasing power of the future returns from securities. If investor-consumers have diverse consumption tastes or face different commodities prices, homogeneity cannot obtain; for, even if they held identical expectations regarding returns the corresponding purchasing powers of returns would nevertheless differ for various investors. This difficulty may be more acute internationally if tastes differ between nations more than they differ within nations. The only separation theorem so far which handles diversity of tastes is that of B. H. Solnik; but one of his assumptions (see Section II) is particularly difficult to accept.

When on the other hand, the market is constrained Pareto optimal *(CPO)*, Nielsen had to introduce additional restrictions in order to preserve *CIV* and to prove that value maximizing firms make *constrained* Pareto optimal production decisions. He assumed essentially that the basis of the linear state-of-nature subspace spanned by existing securities was not modified by the linear transformation (in the case of the *CIV* theorem) or by the firm's decision (in the case of the value-maximization theorem). This means that the cash flows of all pretransformation (resp. predecision) securities must be perfectly correlated with a linear combination of posttransformation (resp. postdecision) securities' cash flows and vice versa. But, just to take an example, this condition is violated as a firm changes its debt-equity ratio even if there are no bankruptcy costs; indeed, since the probability of bankruptcy is usually affected, the state-dependent cash flow pattern of the debt and the equity is modified in a manner which will usually prevent replication of the old pattern by means of linear combinations. I.e., securities have been created or eliminated which are not substitutable with others; this modification of the security-space dimensions implies *discrete changes* in securities values; under such circumstances *CIV* is violated and we do not know what the correspondence is between value maximization and Pareto optimality.

The theory of the firm in the open economy which we shall now review postulates value-maximization as the only practicable objective. At the present state of our knowledge, it must therefore rest on an assumption of *UPO* of the capital market or on the assumption that whatever is done to securities' cash flows preserves the basis of the linear state space.

[3]The value-maximization theorem was established in the absence of taxes and bankruptcy costs. Because the latter involve complex problems of uncompensated social redistribution, it is not at all clear that the theorem is applicable in their presence. This remark casts a shadow over the discussion of the financing decision in Section III.

[4]See D. Cass and Joseph Stiglitz, S. A. Ross, F. Black

II. The Objective of the Firm in the Open Economy

Maximization of the value of the firm's securities frequently calls for the use of a model which gives securities' market values as a function of the probability distribution of their cash flows. In the international (i.e., multi-currency) sphere,[5] two such International Asset Pricing Models (*IAPM*'s) are currently in existence: that of Solnik and that of F. Grauer-Robert Litzenberger and R. Stehle *(GLS)*.

The *GLS* model depends on the assumptions that all goods are traded (so that they are priced at parity everywhere) and that all investors have the same consumption preferences. If a real risk-free asset exists, these assumptions permit a straightforward extension of the closed-economy capital asset pricing model to the world as a whole. Exchange rates in this model can fluctuate only as a result of pure monetary factors and not as a result of relative price fluctuations. The model conveys therefore solely the impact of what can be called "monetary exchange risk."

Solnik, by contrast, assumes that in each country the rate of inflation is zero (or non-random). Contrary to *GLS*'s allegations Solnik's assumption does not imply any money illusion or any inconsistency with the assumed randomness of exchanges rates. So long as consumption preferences differ from country to country, exchange rates can fluctuate if the relative prices of national consumption baskets fluctuate. Furthermore, some goods may not be traded internationally and may therefore not be priced at parity everywhere. Solnik's model conveys therefore the impact of that component of exchange risk which results from relative-price uncertainty, a component which could perhaps be called "real exchange risk." Unfortunately, Solnik's mathematical derivations rely on the assumption that nominal securities' returns are stochastically independent of exchange rates. While this assumption is introduced only for technical reasons, it is nonetheless disturbing economically: since exchange rates in this model represent relative prices, it is unreasonable to assume that they will not affect the value of the firms' outputs and therefore the nominal returns.

To date no *IAPM* exists which incorporates both monetary and real exchange risks, and yet remains usable. Indeed the most general model with heterogeneous consumption tastes would necessarily involve the investors' risk aversions individually,[6] and, since the latter are almost impossible to measure, would be practically useless. The quest today is for a special case which allows treatment of both types of exchange risk, which yields a usable model and which nonetheless remains relevant.

The impact of exchange risk on the firm's valuation maximand and on its equilibrium decisions remains to be worked out precisely. With the help of the above two models, we can nevertheless anticipate the results. Real exchange risk is obviously relevant. its increase should cause a fall in welfare, a rise of current consumption, a decrease of inputs and therefore a fall in firms' values. Monetary exchange risk will frequently be relevant but only because of the existence of nominally defined financial claims. money, debt, etc., . . .; its increase amounts to a redefinition of these securities much as a switch of their denomination would. But a redefinition of securities, although it always leads to a rebalancing of investors' portfolios, may or may not affect consumption and production decisions. If the capital market is and remains *UPO*, if there are no taxes or bankruptcy costs and if money is not issued exclusively by governments, the redefinition will have no real effect and welfare will be un-

[5]The multiplicity of currencies is not the only additional problem to be tackled by international financial theory. Partial segmentation of capital markets is another, see R. Banz. The presence of large oligopolistic firms such as multinational corporations is yet another, see the literature on Direct Investment which unfortunately draws very little on financial theory (see G. C. Hufbauer and the references therein).

[6]See Mark Rubinstein.

affected. If the market is or becomes *CPO*, the redefinition will have unpredictable effects. If there are taxes or bankruptcy costs, the firm will change its financing decision and perhaps consequently also its production decision, as a result of the change in bankruptcy risk.

III. Long-Term Decisions: Financing and Capital Budgeting

We proceed to examine the financing, capital-budgeting and foreign-acquisition decisions of the international firm.[7] In a unified and perfect world capital market where *CIV* can be assumed to hold, the financing, the currency-denomination of debt and foreign acquisition decisions, as we saw, are irrelevant. In the absence of currency-transaction costs, there will also be no optimal choice of currencies in the liquid asset portfolio. The firm need not do what investors can do for themselves. However, when these decisions may change cash flows, as they will in the presence of bankruptcy costs or taxes, optimal decision rules can be derived. To illustrate, let us discuss the financing decision: other decisions follow identical principles. The total value of the firm is decomposed into three terms:

$$V = \Sigma_i V_i = V_o + \Sigma_i t_i D_i - BC(\ldots \Sigma D_i, \Sigma V_i \ldots) \tag{1}$$

Where:

V_i = the total market value of the ith, full-owned corporate subentity (subsidiary or parent) *including* current cash flow.

D_i = face value of the debt issued by the ith subunit.

t_i = effective tax-rate applicable to the debt of the ith subentity.

BC = the present market value of the anticipated bankruptcy costs.

In equation (1) V_o is the market value of the whole firm when no debt is issued. The second term represents the value of the tax savings owing to the deductibility of interest payments everywhere. The third, the market value of the many kinds of costs accruing in the event of bankruptcy, is therefore a function of many variables of which two are especially important: ΣD_i and ΣV_i.[8] ΣD_i determines the promised interest and sinking fund payments which constitute a cash drain, while ΣV_i is a proxy for the ability to meet fixed financial changes. If subsidiaries are not allowed to go bankrupt independently, bankruptcy will occur when the market value of the whole corporate entity's stock reaches zero.[9]

Once the problem is thus formulated, the optimal financing decision can be found. As borrowing increases the value of bankruptcy costs increases slowly at first and then rapidly as the probability of bankruptcy increases. Their difference should reach an optimum. When the value maximand is homogeneous of degree one, the optimum will be reached at some optimal debt/value ratio, $(\Sigma D_i/V)^*$. Individually, $(D_i/V_i)^*$ should be greater the higher the sub-

[7] With an *IAPM* in hand, it will no longer be necessary to assert a utility function for the firm as Agnar Sandmo did for the closed economy, Peter Kenen and Wilfried Ethier for the trading firm and Martin Prachowny for the *MNC*. In a recent paper David Baron derived unanimously preferred decision rules for an international firm. The adoption of this more general objective was costly analytically. Had he assumed price-taking behavior, value maximization would have yielded identical decisions while being simpler to use. Moreover in his model the only source of uncertainty was the exchange rate and there were no taxes or bankruptcy costs.

[8] These costs include lawyers fees, reorganization costs, lost production and/or market position and capital losses if illiquid assets must be sold at a discount. Expropriation costs could also be added to this term.

[9] Bankruptcy will occur at the end of a period if that period's cash throw-offs plus liquid assets on hand are so adverse as to make it unprofitable for the stockholders to prevent bankruptcy by investing further to secure the future cash flow stream. We assume for the multi-unit corporation that bankruptcy involves the whole. Bankruptcy of parts involve contractual nuances which we cannot discuss here.

sidiary's effective tax rate. An open question is the effect of constraints which prevent $(D_i/V_i) > 1$ if this were otherwise optimal.

Subsidiaries need not borrow in their own local currencies. The borrowing decision therefore possesses two additional dimensions which become relevant in the presence of possibly costly bankruptcy: its *c-d* and the place where debt is issued, i.e., the nationality of the bond-holders. However, no models exist that determine the *c-d* mix for long term debt which would (partially) minimize the present value of bankruptcy costs. As in the choice of the optimal *c-d* mix of short-term instruments, discussed below, it is a portfolio problem the solution to which is probably *not* the mix that best matches the *c-d* mix of assets. Finally, the nationality of bondholders should be irrelevant in a unified capital market unless it affects the probability of expropriation. If governments are more likely to expropriate firms whose bonds are held by their nationals, this probability may be a function of the source of the debt: the sourcing decision then affects firm value.

We now turn to the capital budgeting decision and leave the financial decision, having determined an optimal debt ratio. To avoid the problem of simultaneity between the financing and investment decision which is not resolved even in the one-country setting, we assume without justification that firms invest in projects which do not affect firms' costs or risks of bankruptcy.[10] In that case, the preinvestment optimal debt ratio is not changed and can be used as a target for the financing mix of the project. Investment decision rules can be contrived in two cases. In the first, which requires no knowledge of an *IAPM*, the random, after-tax cash flows from the marginal project, $d\tilde{X}$, are perfectly correlated with a linear combination of cash flows yielded by existing securities, $\tilde{X}_j$, $j = 1, \ldots n$:

$d\tilde{X} = \Sigma_j d\delta_j \tilde{X}_j$, where $d\delta_j$ = an infinitely small weight.

[10] See Nitzan Weiss and the references therein

By *CIV*, the marginal contribution of the project to the value of the firm is equal to the *same* linear combination of the *prices* of existing securities *plus* the value of the tax saving due to the project's debt.

$$dV = \Sigma_j d\delta_j V_j + \Sigma_i t_i dV\left(\frac{D_i}{V}\right)^*$$

Accept the project if $dV \geq dI$ = the present cost of the project; i.e., if

$$\Sigma_j V_j \frac{d\delta_j}{dI} \geq 1 - \Sigma_i t_i \left(\frac{D_i}{V}\right)^*$$

This decision rule can be recast in cost-of-capital terms. Compute the weighted average, ρ of the yields, ρ_j, on existing securities: $\rho = \frac{\Sigma_j d\delta_j \rho_j V_j}{\Sigma_j d\delta_j V_j}$ where $\rho_j = \frac{E(\tilde{X})}{V_j} - 1$ Then, accept if $\frac{E(d\tilde{X})}{dI} \geq (1 + \rho)\left[1 - \sum_i t_i \left(\frac{D_i}{V}\right)^*\right]$

The after-tax, zero-debt, rate of return on the project must be larger than its cost of capital adjusted for taxes. These criteria which generalize Franco Modigliani-Merton Miller are not essentially different from what they would be in the closed economy.[11] Only the target debt ratios have an international dimension. For a given pattern of bankruptcy risks and costs and, therefore, a preset *c-d* decision, these decision criteria will be independent of exchange risk.[12] Clearly, also, firms which have different bankruptcy costs and risks will generally evaluate a given project differently. But it is unlikely, pending further research, that decision criteria will vary systematically across countries. We do not mean to imply, however, that project

[11] Modigliani and Miller restricted their criteria to projects in the firm's risk class, their weighted average cost of capital consequently included only the firm's own debt and equity. Our criteria are not so restricted.

[12] To be sure, changes in exchange rate variances would change the values of all securities and of the market's risk adjustment (the market price of risk). But price taking firms would take these as given

decisions themselves will be independent of locational factors, for project cash flows, $d\tilde{X}$, will generally vary by location. Equally, these cash flows will be affected by exchange risk to the extent that projects employ nominal balances.

In the second case, the project is not perfectly correlated with a linear combination of existing securities. Assuming that sufficient conditions for *UPO* are met, investment decision criteria like those above can be derived from a relevant *IAPM*, of the type discussed in Section II, following well known procedures, initiated by Robert Hamada. The additional problems with this approach are familiar: the market adjustment for risk varies with the assumed characteristics of investors' utility functions; and the projects risk characteristics must be measured prospectively, in practice a difficult task.

To complete the discussion of long-term decisions let us move briefly to a world of completely segmented capital markets: investors do not invest abroad at all. If firms can bridge the gulf between national segments, *CIV* no longer obtains, and there appears a one-to-one correspondence between the market which is segmented and the decision which becomes relevant: if the bond market is segmented, the choice of the locus of borrowing is relevant, if the stock market is, the foreign acquisition decision becomes relevant. Let us review the latter.

Assume that each national market, separately, is *UPO* and that within each segment exchange and inflation risks are perceived homogeneously. Assume further, following Adler-Dumas and W. Y. Lee-K. S. Sachdeva that the firms of only one country are permitted to invest abroad, but behave as price-takers in both markets. Value maximization will be Pareto-optimal from the point of view of home market investors who take the prices of foreign securities as given but not from that of foreigners. The optimal aggregate acquisition decision can then be derived from a Capital Asset Pricing Model *(CAPM)* into which the postacquisition cash flows of home firms are substituted. It satisfies three conditions: (a) foreign securities are priced as if they were valued by the home *CAPM;* (b) foreigners' excess demand at those prices is filled by the aggregate acquisition decision; and (c) the allocation of the aggregate among individual firms is indeterminate. Because the home market is already capable of an optimal allocation of risk-bearing, foreign securities provide no diversification advantage.[13] No results are presently available for the more general case which would permit foreign acquisitions by firms everywhere.

IV. The Short-Term Decisions

The short-term decisions of the firm are the area of overlap between the theory of the firm and monetary theory, since they involve among other things the firm's demand for cash and other liquid assets and liabilities. But the relationship of these decisions to value maximization is not yet clear in the closed economy much less the open economy. When attempting to maximize the market value of its securities with respect to short-term decisions, the firm should presumably take three categories of costs into consideration: opportunity costs, transaction costs and bankruptcy costs. The opportunity costs arise out of the fact that short-term assets such as cash have a rate of return lower than fixed assets; keeping a larger proportion of short-term assets therefore reduces the long-term profitability of the firm and thereby also, paradoxically, increases the probability of future bankruptcy.[14] Transaction costs have

[13]In this case value maximizing investment decisions by both the parent and the acquired subsidiary will be Pareto optimal for home-country investors (though not for foreigners). Adler examined the case where the home market is *UPO* but, conversely, the parent does not act as a price taker in the foreign market. The parent has an optimal acquisition to make for home country investors. However, value maximization by the foreign subsidiary with respect to its own decisions ceases to be Pareto-optimal for investors in either capital market.

[14]This point is made by Weiss in his Columbia University, Ph.D., dissertation in progress.

been neglected in the context of long-term decisions; to the extent that short-term decisions involve more frequent transactions, however, it is possible that these costs will no longer be negligible Finally bankruptcy costs have already been described; one of these is the loss if assets must be liquidated at a discount: short-term assets usually carry a smaller liquidation discount (they are more "liquid") than do long-term assets. Hence increasing the proportion of short-term assets reduces bankruptcy costs Following from these considerations, an optimal short-term decision generally exists.

Unfortunately no model yet exists which maximizes firm value taking all three categories of costs into account. To date there are two strands of models each of which accommodates, wholly or in part, pairs of cost items. In the first strand, opportunity costs (not, however, explicitly linked to firms' costs of capital as defined above) plus transactions costs are minimized. These are usually inventory theoretic models, related to monetary theory in the spirit of the pioneering work of William Baumol and Tobin. The approach has been extended by Sho Tsiang to incorporate uncertainty but has not been applied internationally although it could fruitfully be used to derive propositions regarding the degree of substitutability of currencies' balances with each other and with short-term instruments of various denominations, depending on the level of transactions costs and the degree of fluctuation of exchange rates A model which seeks to minimize the proportional transactions costs of maneuvering liquid assets internationally is that of D P. Rutenberg; it is designed to optimize not the level of liquid assets in each location, but rather their flows among corporate subunits. The transportation algorithm is used and uncertainty is not taken into account.

The models of the second strand minimize bankruptcy costs and opportunity costs in a portfolio choice approach. Since this approach is characteristic of financial theory and has already found tentative international application, we pursue it below.

Consider first the short-term decision problem in the closed economy. Assuming for simplicity that investors are risk-neutral, the impact of bankruptcy costs on the value of the firm is simply equal to the probability of bankruptcy multiplied by the expected costs conditional on the event of bankruptcy. Suppose that the dominant component of bankruptcy cost is the discount contingent upon the liquidation of assets; each asset is then characterized by its rate of discount and the total rate on the whole portfolio of the firm's assets and liabilities is equal to a weighted average of the individual discount rates.

Following the definition of bankruptcy in footnote 9, there exists in each period a required minimum rate of return $\tilde{R}^*$ on the firm's total portfolio of assets and liabilities, below which the firm is bankrupt If these rates of return are all normally distributed, the probability of bankruptcy is a function only of the coefficient of variation (expected value over standard deviation) of the firm portfolio's excess rate of return over the required rate. In that case we can write the firm's portfolio choice in the following way.[15]

$$\text{Maximize } E(R_p) - C \times \emptyset \cdot \left[\frac{-E(\tilde{R}_p) + E(\tilde{R}^*)}{\sqrt{\{\sigma^2(\tilde{R}_p) + \sigma^2(\tilde{R}^*) - 2\operatorname{cov}(\tilde{R}_p, \tilde{R}^*)\}}} \right]$$

Where. $E(\tilde{R}_p) = \Sigma_j x_j E(\tilde{R}_j)$

$\sigma^2(\tilde{R}_p) = \{\Sigma_i \Sigma_j x_i x_j \operatorname{cov}(\tilde{R}_i, \tilde{R}_j)\}$

$C = \Sigma_i x_i b_i =$ bankruptcy cost

$b_i =$ liquidation discount on asset or liability i

$x_i =$ relative weight of the asset or liability in the portfolio; $x_i > 0$ for an asset and $x_i < 0$ for a liability

$\emptyset =$ area under the normal curve

[15] This objective is applicable only when investors are risk neutral In the case of risk aversion, one should subtract a risk penalty on the portfolio return R_p and add back a risk premium on the costs of bankruptcy in order to reconstitute the value-maximization objective

Besides the obvious constraint $\Sigma_i x_i = 1$, there may be others. For instance, there is a minimum required amount of cash to be kept for transactions purposes. Accounts receivable and accounts payable may be related rigidly to sales if the firm is forced by its competitors to adopt a specific credit or payment period. Finally, the x_i's corresponding to long-term assets and long-term debt are not objects of choice as they have been determined previously on other grounds;[16] their presence nevertheless would influence the short-term portfolio choice.[17] The above problem can be solved by repetitive quadratic programming in the manner of Markowitz and by varying the value of C whose definition serves as a constraint. The value of C and the corresponding portfolio which maximizes the objective is easily selected

Extension of this framework to the multicurrency world is promising. Section II indicated that exchange risk would affect the probability of bankruptcy and therefore in the presence of bankruptcy costs the value of a firm which holds nominal financial claims: there will be an optimal c-d mix in this situation for short as well as long-term financial decisions. One can therefore hope in the future to link studies such as B. A Lietaer's, which arbitrarily seeks to minimize losses due to exchange risk subject to rate of return constraints, via the bankruptcy cost nexus to the valuation objective. However, the quadratic programming approach cannot, rigorously speaking, be applied directly to the multicurrency case because the assumption of universal normality of returns breaks down. If the returns are normal in one currency, they are not in another because of multiplication by the random exchange rate

[16]The level of long-term debt presumably cannot be altered as frequently as the portfolio of short-term assets and short-term liabilities. This as well as the tax advantage of the debt would be the chief differences between the above program and that which determined the long-term debt; but ideally the two should be solved simultaneously cf Weiss

[17]See David Mayers's treatment of the "non-traded-asset" problem The long-term assets and long-term debt would act as a nontraded-asset with respect to the choice of the short-term portfolio

Lietaer while not addressing the valuation problem side-stepped this difficulty by means of an approximation. He approximated the correct formula for translating returns, shown in the lefthand side of the following equation, by the one shown on the righthand side:

$$\frac{(1+\tilde{R})\tilde{S}}{S_0} - 1 \sim \tilde{R} + \frac{\tilde{S} - S_0}{S_0}$$

where: $\tilde{R}$ = foreign-currency return

$\tilde{S}$, S_0 = future and current exchange rate

This approximation would be legitimate when applied to nonstochastic variables; when applied to random ones, however, it can be shown that it leads to a portfolio choice which is approximately optimal only when the rates of return and the exchange rates are approximately independent. If that condition is verified, it suffices to postulate normality of $\tilde{R}$ and $\tilde{S}$ to obtain normality of converted returns and the quadratic programming formulation can then be used.[18] One major disturbing aspect of the approximation, however, is that it leads to somewhat different portfolio choices depending on the currency in which the returns are measured.

Any prospective solution to the above portfolio choice problem, whether exactly or approximately specified, will undoubtedly call for diversification of the currency portfolio. This conclusion runs counter to the practice of some corporate treasurers who hedge returns denominated in currencies other than the parent's in an attempt to reduce their "exchange-risk exposure" (cf. Kenen). In fact, when structuring such portfolio-choice models, a change of variables is sometimes convenient. This amounts to defining a variable which is the sum of the firms balance-sheet weights corresponding to assets and liabilities denominated

[18]All the short-term decisions would fall out of this program In particular the decision to enter forward contracts, and the choice of denomination of all short-term assets and liabilities to the extent that they can be chosen, would thus be determined The optimal denomination of accounts receivable leads to the choice of a currency for invoicing

in the same currency. This is exactly what accountants refer to as exposure. There is no reason, however, to institute the changed variable as a target to be minimized for some currencies.

This argument extends further. Lietaer introduces a totally different definition of exposure which seems more useful than the accountants' in the case where $\tilde{R}$ is not independent of $\tilde{S}$. Exposure may intuitively be identified with the partial regression coefficient, b_{ij}, of $\tilde{S}_i(1 + \tilde{R}_{ij})$ with respect to $\tilde{S}_i$, where i subscripts currencies and j, asset returns paid out in a given currency. If $b_{ij} = 0$, the returns are not exposed; if $b_{ij} = 1$, they are totally exposed. The firms total exposure is a weighted average of the individual exposures While this notion can serve as a fine descriptive tool, it is in no way implied that it represents a quantity to be reduced. Furthermore in multiperiod models exposure will remain a concept describing cash returns of each period separately It is usually impossible to define an aggregate over several periods which could be called the exposure of an asset or liability taken as a whole.

REFERENCES

Michael Adler, "The Cost of Capital and Valuation of a Two Country Firm," *J. Finance*, Mar. 1974

______ **and Bernard Dumas,** "Optimal Internal Acquisitions," *J. Finance*, Mar. 1975.

______ **and Guy V. G. Stevens,** "The Trade Effects of Direct Investment," *J Finance*, May 1974, *29*, 655–76.

R. Banz, "Capital Asset Pricing in Partially Segmented Markets," University of Chicago, Graduate School of Business, Jan. 1976

David P. Baron, "Flexible Exchange Rates, Forward Markets, and the Level of Trade," *Amer. Econ. Rev.*, June 1976, *66*, 253–67.

William J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretical Approach," *Quart. J. Econ.*, Nov. 1952, *26*, 545–56.

F. Black, "Capital Market Equilibrium with Restricted Borrowing," *J. Business*, Mar. 1973.

David Cass and Joseph Stiglitz, "The Structure of Investor Preferences and Assets Returns, and Separability in Portfolio Selection: A Contribution to the Pure Theory of Mutual Funds," *J. Econ Theory*, 1970, *2*, 122–60

Wilfrid Ethier, "International Trade and the Forward Exchange Market," *Amer. Econ. Rev.*, June 1973, *63*, 494–503.

E. F. Fama and Merton H. Miller, *The Theory of Finance*, New York 1972.

F. Grauer, Robert H. Litzenberger and R. Stehle, "Sharing Rules and Equilibrium in an International Capital Market Under Uncertainty," *J. Financial Econ.*, June 1976, *3*, 233–56.

Robert S. Hamada, "Portfolio Analysis, Market Equilibrium and Corporation Finance," *J. Finance*, Mar. 1969, *24*, 13–32.

T. Horst, "The Theory of the Multinational Firm. Optimal Behavior Under Different Tariff and Tax Rates," *J. Polit Econ*, Sept.-Oct. 1971, 1059–72.

G. C. Hufbauer, "The Multinational Corporation and Direct Investment," in P. B Kenen, ed., *International Trade and Finance*, Cambridge 1975, 253–320.

Peter B. Kenen, "Trade, Speculation and the Forward Exchange Rate," in Baldwin, et al., *Trade Growth and the Balance of Payments*, Chicago 1966.

W. Y. Lee and K. S. Sachdeva, "The Role of the Multinational Firm in the Integration of Segmented Capital Markets," Indiana University, School of Business, July 1976

B. A. Lietaer, *Financial Management of Foreign Exchange*, Cambridge, Mass 1971.

H. Markowitz, *Portfolio Analysis*, New York 1959

David Mayers, "Non Marketable Assets and Capital Market Equilibrium Under Uncertainty," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets*, New York 1971, 223–48.

Franco Modigliani and Merton H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

Neils C. Nielsen, "The Firm as an Intermediary Between Consumers and Production Functions Under Uncertainty," Ph.D Dissertation, Stanford University 1974

Martin F. J. Prachowny, "Direct Investment and the Balance of Payments of the United States: A Portfolio Approach," in F. Machlup, et al., *International Mobility and Movement of Capital,* New York 1972

S. A. Ross, "Mutual Fund Separation in Financial Theory—The Separation Distributions," University of Pennsylvania, R. White Center 1976.

Mark Rubinstein, "An Aggregation Theorem for Securities Markets," *J Financial Econ.*, Sept 1974, *1*, 225–44

D. P. Rutenberg, "Maneuvering Liquid Assets in a Multi-National Company. Formulation and Deterministic Solution Procedures," *Management Science,* June 1970, *17*.

Agnar Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

J. H. Scott, "On the Theory of Conglomerate Mergers," Columbia Graduate School of Business, Working Paper No. 135, June 1976.

B. H. Solnik, "An Equilibrium Model of the International Capital Market," *J. Econ. Theory,* Aug. 1974, *8*, 500–24.

Guy V. G. Stevens, "Fixed Investment Expenditures of Foreign Manufacturing Affiliates of U.S. Firms: Theoretical Models and Empirical Evidence," *Yale Economic Essays,* Spring 1969.

James Tobin, "The Interest-Elasticity of Transactions Demand for Cash," *Rev. Econ. Statist.*, Aug. 1956, *38*, 241–47.

Sho C. Tsiang, "The Precautionary Demand for Money: An Inventory Theoretical Analysis," *J. Polit. Econ.*, Jan.-Feb. 1969, *77*, 99–117.

Nitzan Weiss, "A Simultaneous Solution to the Real and Financial Decisions of The Firm Under Uncertainty: An Integration of the Neoclassical Theory of the Firm and Finance Theory," Columbia University, Graduate School of Business, 1976.

Modeling the Interdependence of National Money and Capital Markets

By Dale W. Henderson*

During the last decade there has been a thoroughgoing reworking of monetary theory for open economies using a portfolio balance approach. According to this approach, equilibrium in international financial markets occurs when the available stocks of national moneys and other financial instruments are equal to the stock demands for these assets based on current wealth, and wealth accumulation continues only until current wealth is equal to desired wealth.

In this paper some of the more important results that can be obtained using open-economy portfolio balance models are discussed and, in some cases, contrasted with results from earlier models. A model of a single open economy, which is illustrative of the type of model currently in use, is described.[1] Using this model as an expositional device, the effects of policy actions on endogenous variables are traced out over three time horizons: 1) a short run in which only financial variables adjust, 2) a short run in which both financial and nonfinancial variables adjust, and 3) a long-run stationary state in which the stock of wealth adjusts to its desired level.

I. Financial Markets in the Short Run

In short-run, portfolio balance models of international financial markets, price levels, incomes, and initial asset holdings are taken as exogenous, and financial markets are equilibrated by rapid adjustments in interest rates, the financial asset holdings of wealth holders and central banks, and, under flexible exchange rates, the exchange rate.[2] As a concrete example, consider an open economy in which residents hold home money, home securities, and foreign securities. The fraction of the nominal value of wealth which home residents want to hold in each asset depends on the home interest rate and the foreign interest rate adjusted for the expected rate of depreciation of the home currency price of foreign currency. The expected rate of depreciation of the exchange rate is assumed to be zero under fixed exchange rates; under flexible exchange rates it is taken to be a decreasing function of the gap between the current exchange rate and a constant "long-run equilibrium" level of the exchange rate.[3] The three assets held are regarded as strict gross substitutes. Foreigners' demand for home securities measured in foreign currency depends on the home interest rate and on the foreign interest rate adjusted for the expected depreciation of the home currency. It is assumed that the foreign authorities peg the foreign interest rate. The home money supply is the liability of the

*Board of Governors of the Federal Reserve System. It is hoped that Richard Berner, Ralph Bryant, Peter Clark, Betty Daniel, Michael Dooley, Lance Girton, Dwight Jaffee, Jurg Niehans, Larry Promisel, Don Roper, Jeffrey Shafer, Edwin Truman, and Janet Yellen find this paper improved as a result of attempts to take account of their helpful comments. The author is responsible for the paper's remaining shortcomings. The analysis and conclusions of this paper should not be interpreted as representing the views of the Board of Governors of the Federal Reserve System or anyone else on its staff.

[1]Attention is confined to the case of a single open economy for simplicity. It is assumed that the foreign policy authorities act to prevent changes in foreign variables which otherwise would affect behavior in the home economy. To study the implications of other possible foreign policy responses, it is necessary to employ two-country or "world" models such as the ones used by Stanley Black, Charles Freedman, Lance Girton and Henderson, and John Hewson and Eisuke Sakakibara. Two-country models can also be used to explore the game theoretic aspects of macroeconomic policy making in open economies.

[2]Examples of models of this type are Black, Freedman, Girton and Henderson, and Hewson and Sakakibara.

[3]That is, for simplicity, it is assumed that exchange rate expectations are regressive or stabilizing. If expectations are extrapolative or destabilizing, a depreciation in the exchange rate increases the expected rate of depreciation. Black, Rudiger Dornbusch (forthcoming), and Pentti Kouri explore the implications of assuming that expectations are rational.

central bank and is matched on the asset side by central bank holdings of domestic securities and of reserve assets like gold, Special Drawing Rights, foreign money or foreign securities. The supply of home securities to the public is equal to the exogenous supply of fixed-nominal-value, variable-interest-rate securities issued by the government of the home country minus the holdings of the central bank.[4] Equilibrium occurs when the stock demands for and supplies of home money and home securities are equal. Two variables, the home interest rate and either the reserve stock of the home central bank or the exchange rate adjust to equilibrate these markets.

A set of results which by now have become well known can be derived from models of the type just described. When the central bank undertakes an open market purchase under fixed exchange rates,[5] the excess demand for home securities and the excess supply of home money cause a decline in the home interest rate and put pressure on the exchange rate to depreciate. The central bank intervenes, purchasing home money with foreign money, and experiences a corresponding loss of reserves. The home country money supply is increased by an amount less than the amount of the open market purchase because only part of the initial excess supply of money is matched by an increase in the demand for money resulting from the lower interest rate; the rest is removed through intervention.

This result is modified in familiar ways under extreme assumptions. If the home country is small in the market for home securities[6] or if home and foreign securities are perfect substitutes, the home interest rate and, therefore, home money demand cannot be changed by an open market purchase, and the home country loses reserves equal to the full amount of the open market purchase.[7] That open market purchases of domestic securities not matched by increases in money demand lead to corresponding reserve losses is one important lesson of the monetary approach to the balance of payments.[8] Decreases in relative economic size and increases in the degree of substitutability short of the extreme cases reduce the effectiveness of a given open market purchase in lowering the domestic interest rate and in raising the money supply and increase the associated loss of reserves.

These results are important for two reasons. First, unless the authorities can affect the domestic interest rate and money supply, they cannot use monetary policy to affect ultimate policy objectives like home prices and output. Second, when the exchange rate is a target of policy and the stock of home country reserves is given, the larger the reserve loss associated with a given open market operation, the less freedom the home country monetary authorities have to pursue an independent monetary policy.

If the exchange rate is allowed to float, an open market purchase by the home central bank puts downward pressure on the interest rate and causes the exchange rate to depreciate. This depreciation stimulates demand for home money and home securities for two reasons. First, there is the conventional expectations effect. A depreciation of the exchange rate reduces the expected rate of depreciation, thereby stimulating demand for home currency assets. Second, there is the less conventional valuation effect. The home currency value of home wealth increases, but this increase accrues completely in the form of a rise in the value of hold-

[4]It is assumed that tax liabilities are not marketable and that the effects of changes in tax liabilities on saving and asset demands are negligible. It is also assumed that the foreign central bank holds no home money or home securities.

[5]Whenever the designation "fixed exchange rates" is used it is assumed that the central bank does not sterilize the home money supply by undertaking open market operations to offset the effects of reserve losses on the home money supply. The effects of sterilization on some of the results are discussed in the footnotes.

[6]The home country is small in the market for home securities if shifts in home demand for home securities or home central bank open market purchases are negligible in size compared to the responsiveness of foreign demand for the home security to changes in the home interest rate.

[7]Note though that if home and foreign securities are perfect substitutes but the home country is not small and if the foreign country does not peg the world interest rate, the home country can lower this interest rate at the cost of some loss of reserves.

[8]For expositions of this approach see Harry Johnson and Michael Mussa.

ings of foreign securities. To rebalance their portfolios domestic residents seek to acquire more domestic assets.[9] Of course, the home money supply rises by the full amount of the open market purchase. The home interest rate falls by more than under fixed rates since exchange rate depreciation adds private demand to the central bank demand for home securities. If the home country is small or if home and foreign securities are perfect substitutes so that the home interest rate equals the foreign interest rate plus the expected rate of depreciation, the home interest rate falls only because of the expectations effect of the exchange rate depreciation, but money demand rises to match the new larger money supply whether or not the expectations effect is at work because of the valuation effect.

Short-run financial models can also be used to analyze the effects of exchange market operations or intervention policy. An exchange market operation, like an open market operation, involves an exchange of assets between the home central bank and the public, but in the case of an exchange market operation, the assets exchanged are denominated in different currencies. In one type of exchange market operation the central bank supplies home money and demands foreign securities.[10] This action depreciates the exchange rate, and, since this depreciation raises the demand for home securities through the expectations and valuation effects, causes the interest rate to fall. The exchange rate depreciation is larger, and the interest rate decline is smaller than in the case of an open market purchase of the same size because the exchange market operation does not involve demand by the central bank for home securities. If home and foreign securities are perfect substitutes, an exchange of home money for foreign securities has the same effect as an open market operation.

II. Adding the Goods Market in the Short Run

The standard macroeconomic model of full short-run equilibrium in the open economy in use in the mid-1970's differs from the one in use in the mid-1960's in an essential way.[11] For financial markets it incorporates a stock equilibrium formulation like the one sketched out in the last section. An implication of this formulation is that the basic set of equilibrium conditions does not include an equation for the time derivative of home country reserves; that is, the equation which was referred to in the 1960's as "the foreign exchange market" equation and which often embodied a flow formulation for the capital account no longer plays a central role. This difference implies that some of the results of the earlier models no longer hold.

Consider an example which is typical of some of the recent formulations. The home country produces a single traded good which is an imperfect substitute for a single foreign good in the consumption of both countries. The money wage rate is fixed in the short run so that production of the home good is an increasing function of its price. There is no capital stock or investment. Lump-sum taxes are equal to government spending plus interest payments on the government debt. The expected rate of inflation of the prices of goods is zero. The foreign authorities fix the foreign currency price of the single foreign good, and all real variables are

[9] It is assumed that home country residents have net assets denominated in foreign currency. Under this assumption a depreciation may increase home demand for home currency assets even if exchange rate expectations are extrapolative. Of course, if home country residents have net liabilities denominated in foreign currency, a depreciation may reduce home demand for home currency assets even if expectations are regressive. For futher discussion of the valuation effect, see Girton and Henderson.

A depreciation also increases the home currency value of foreigners' demand for home securities.

[10] It does not matter whether foreign securities or foreign money are purchased since the foreign central bank keeps the foreign interest rate constant. For a further discussion of exchange market operations, see Girton and Henderson.

[11] Examples of the models of the 1970's are those of Russell Boyer, William Branson, and Dornbusch (1975). Representative of the models of the 1960's are the ones employed by J. Marcus Fleming and Robert Mundell. Mundell's Chapter 18 contains a model which is consistent with the portfolio balance approach.

measured in terms of the foreign good.

Equilibrium in the home goods market requires that saving equal the current account surplus. Home saving depends positively on real disposable income, the home interest rate, and the foreign interest rate adjusted for the expected rate of depreciation of the home currency and negatively on real wealth. Disposable income is equal to the production of the home good minus lump-sum taxes plus interest receipts from holdings of home and foreign securities. The current account surplus is equal to the trade account surplus plus net interest receipts. The trade account surplus depends negatively on home private spending (disposable income minus saving) and on the relative price of the home good.

In order to take account of the transactions demand for money, it is assumed that the fraction of home nominal wealth held in home money depends positively on the ratio of the value of home production to nominal wealth and that the fractions of home nominal wealth held in home and foreign securities depend negatively on this ratio.

In the illustrative model, three equilibrium conditions, the home goods market condition, the home money market condition, and the home securities market condition determine three variables, the home price level (and output), the home interest rate, and either the home country's stock of reserves or the exchange rate.[12] In contrast, in the models of the 1960's three equilibrium conditions, the home goods market condition, the home money market condition, and the foreign exchange market condition determine three variables, the home price level (and output), the home interest rate, and either the time derivative of reserves or the exchange rate. The typical foreign exchange market equation states that the time derivative of reserves is equal to the current account surplus plus net capital inflow which is assumed to depend positively on the home interest rate and negatively on the foreign interest rate adjusted for the expected depreciation of the home currency.[13]

In the newer models with the exchange rate fixed, an open market purchase causes the interest rate to fall and the price level to rise, but may lead to either a decrease or an increase in central bank reserves. Excess demand for home securities causes the interest rate to fall. Excess demand develops in the goods market, so the home country price level must rise. The home country loses reserves if the fall in the home interest rate and the rise in the home price level required to re-equilibrate the securities market and the goods market cause money demand to rise by less than the original increase in excess supply due to the open market purchase. While an open market purchase definitely causes a loss of reserves in short-run financial models, the possibility that the home country may gain reserves arises in models with an endogenous

[12] Boyer, Branson, and Dornbusch assume that both traded and nontraded securities are held in home country portfolios, and Boyer and Dornbusch assume that the home economy produces both traded and nontraded goods. The interest rate on traded bonds and the price of traded goods are assumed to be set in the world economy. The differences between the model of this paper and those with nontraded securities and goods are not crucial. What is important in all the models is that, except in extreme cases, it is possible, at least in the short run, for policy actions to change an interest rate and a relative price relevant for home decision making.

[13] The representative model of the 1970's described in this paper is a continuous time model with instantaneous adjustment in financial markets. Without sterilization the stock of reserves is determined recursively, but with sterilization the levels of the three endogenous variables are jointly determined. Neil Wallace and Branson have shown how the time derivative of reserves can be determined in this kind of model, but even with sterilization the time derivative of reserves does not affect the levels of the endogenous variables.

A foreign exchange market equation which is somewhat similar to the ones used in the 1960's but which takes account of portfolio balance considerations can be employed as an equilibrium condition in the determination of the levels of endogenous variables in at least two kinds of models: 1) continuous time models with gradual adjustment in financial markets due to adjustment costs and 2) period models with instantaneous adjustment in financial markets in which portfolios are balanced taking into account the additions to wealth which accrue during the period and in which the previous period's values of the endogenous variables are specified.

price level because of the transactions demand for money.[14]

Under flexible exchange rates, an open market purchase causes the interest rate to fall and the price level to rise, but may lead to either a depreciation or appreciation of the exchange rate. The exchange rate depreciates under the same conditions which lead to a loss of reserves under fixed exchange rates. While an open market purchase definitely causes the exchange rate to depreciate in short-run financial models, the possibility that the exchange rate may appreciate arises in models with an endogenous price level because of the transactions demand for money. An open market operation causes the price level (and output) to rise by more or less under flexible exchange rates than under fixed exchange rates depending on whether the exchange rate depreciates or appreciates.[15] In contrast, in the models of the 1960's an open market purchase definitely causes the exchange rate to depreciate because the associated interest rate decline and price level increase both lead to an excess demand for foreign exchange, so the price level rises more under flexible exchange rates.

The portfolio balance formulation of the financial sector of the open economy also has implications for the analysis of the effects of fiscal policy actions. Consider the effects of a balanced-budget increase in government spending under fixed exchange rates. This action reduces disposable income and, therefore, saving; it also reduces private spending and, therefore, improves the current account. These changes give rise to excess demand in the goods market, so the price level rises. This price rise creates an excess supply of bonds and causes the interest rate to rise. The price level and interest rate increases which re-equilibrate the goods and securites markets must lead to an excess demand for money, so the home country definitely gains reserves as a result of the fiscal policy action.[16] This result follows from the assumptions above regarding asset demands; a given price increase raises the excess demand for money by more than it reduces the excess demand for home securities, and a given rise in the interest rate reduces the excess demand for money by less than it increases the excess demand for bonds.

Under flexible exchange rates the price rise caused by the excess demand for goods leads to an increase in the excess demand for money and an excess supply of securities. These financial market disequilibria are removed by an appreciation of the exchange rate and a rise in the rate of interest. Because of the appreciation of the exchange rate, fiscal policy is less effective under flexible exchange rates than under fixed exchange rates. In contrast, in the models of the 1960's a balanced budget expansion may cause the exchange rate to either appreciate or depreciate because the associated interest rate increase and price level increase may lead to ei-

[14] More specifically, the possibility that the home country may gain reserves arises because increases in the price level reduce home country demand for foreign securities. If increases in money demand for transactions purposes are exactly matched by decreases in demand for foreign securities, the home country must lose reserves since the fall in the home interest rate and the rise in the home price level required to re-equilibrate the securities market and the goods market must cause money demand to rise by less than the original increase in excess supply.

[15] Throughout this paper it is assumed that a depreciation of the exchange rate creates excess demand in the goods market. This effect is not a necessary consequence of the behavioral assumptions made above, and some results are altered if a depreciation creates enough excess supply in the goods market. A depreciation tends to raise the excess demand for goods because it 1) reduces the noninterest component of real disposable income, 2) reduces the relative price of the home good, and 3) reduces saving by causing a decline in the expected rate of depreciation. it tends to lower excess demand because it increases saving by causing a decline in real wealth. The effects of a depreciation on the interest component of real disposable income and on the service account are ambiguous.

[16] If the home monetary authorities sterilize the money supply under fixed exchange rates, monetary policy may be less effective under flexible exchange rates even if the exchange rate depreciates. This result arises because depreciation creates excess demand in the money market as well as in the goods market, so the equilibrium interest rate must be higher under flexible exchange rates than under fixed exchange rates with sterilization.

ther an excess supply of or excess demand for foreign exchange, so fiscal policy may be less or more effective under flexible exchange rates.

If the home country is small in the market for its own good,[17] or if the home good is a perfect substitute in consumption for the foreign good, an open market purchase has no effect on the price level under fixed exchange rates, but leads to an increase in the price level under flexible exchange rates since the exchange rate depreciates. Given either of these extreme assumptions, a balanced budget expansion of government spending has no effect on the price level under either exchange rate regime.

It is interesting to re-examine two standard propositions from the theory of economic policy derived from the models of the 1960's in the context of the newer models. In contrast to the result obtained in the models of the 1960's, in the more recent models under fixed exchange rates, fiscal policy has a comparative advantage over monetary policy in achieving external balance, while monetary policy has a comparative advantage over fiscal policy in achieving internal balance. That is, it takes a smaller contractionary fiscal policy action to offset a given open market purchase to keep the reserve stock constant than to keep the price level (and output) constant.[18] The question of comparative advantage arises only when expansionary monetary policy leads to an increase in the reserve stock. Suppose the monetary authority undertakes an open market purchase In the absence of fiscal policy the interest rate would fall, the price level would rise, and, by assumption, there would be an excess demand for home money. In order to keep the reserve stock constant following the open market purchase, the fiscal authority must contract in order to reinforce the interest rate decline and restrain the price level increase so that the excess demand for money is removed while the home securities market remains in equilibrium. At the new reserves-constant equilibrium the price level must be higher than it was before any policy was undertaken; otherwise the home securities and money markets cannot clear simultaneously. Thus, an even more contractionary fiscal policy would be required to keep the price level (and output) constant.

In the newer models it is still true that monetary policy is relatively more effective when compared to fiscal policy in stabilizing the price level (and output) under flexible exchange rates than under fixed exchange rates. Suppose the monetary authority undertakes an open market purchase. To stabilize the price level under fixed exchange rates, the fiscal authority must reduce government spending and taxes by enough to assure that the interest rate falls to a level compatible with the smaller stock of home securities at the old price level. At the old price level with a decrease in the interest rate sufficient to clear the home securities market there is still an excess supply of money, so under fixed exchange rates the country loses reserves. Under flexible exchange rates this excess supply of money would lead to a depreciation of the exchange rate and a drop in the interest rate, so the price level would have to be higher in the new equilibrium than it was before any policy actions were undertaken. Thus, it takes a more contractionary fiscal policy to offset the effects of a given expansionary open market operation on the price level under flexible exchange rates than under fixed exchange rates.

[17]The home country is small in the market for the home good if changes in private or public demand for the home good are negligible relative to the responsiveness of foreign demand for the home good to changes in its relative price.

[18]External balance in the models of the 1960's was defined as the attainment of a desired value for the time derivative of reserves, but in the newer models it seems more natural to define external balance as the attainment of a desired value for the stock of reserves

III. The Stationary State

The last ten years have also seen the development of a coherent description of long-run equilibrium in the open economy, and, since growth in the labor force and technical change have usually been assumed away, the long-run equilibrium typically described has been a sta-

tionary state equilibrium.[19] In such an equilibrium there is no asset accumulation because the actual value of wealth is equal to its desired value; there is no incentive to rearrange portfolios over time since no argument of the asset market equilibrium conditions is changing because no exogenous variable is moving over time; all markets clear because all prices are flexible; and all expectations are fulfilled. The properties of long-run models have a bearing on a number of important questions in international monetary economics. Answers to some of these questions are implied by the existence of a stationary state equilibrium. Answers to others depend upon whether or not the stationary state model exhibits neutrality with respect to certain variables.

For the illustrative model under fixed exchange rates, stationary state equilibrium requires that saving, the current account, the excess demand for home securities, and the excess demand for money all equal zero. These requirements can, in general, be met since there are four endogenous variables, the price level, the interest rate, the stock of real wealth, and the stock of reserves. Given that the money wage rate is flexible, there is full employment, and output is no longer a function of the price level. Given that real wealth and the money wage rate do not change once the stationary state equilibrium is reached,[21] the time derivative of reserves is zero.[22]

Answers to three questions follow immediately if the requirements for stationary state equilibrium are met. First, the model is an "equilibrium system" from the point of view of balance of payments adjustment since the time derivative of reserves is zero in the long run. Put in another, more provocative, way, any exchange rate is an equilibrium exchange rate. The catch, of course, is that the restoration of stationary state equilibrium after an exogenous shock may require large reserve losses, so the adjustment process may not be permitted to follow its full course. Second, a devaluation has no long-run effect on the current account or the time derivative of reserves. In the illustrative model described above, a devaluation does cause a change in the trade account and an offsetting change in the service account. Third, an attempt to increase the money supply holding the nominal supply of securities constant results in an offsetting loss of reserves and leads to no other changes no matter what the degree of substitutability between home and foreign securities.[23] The saving, current account, and securities market equilibrium conditions are not disturbed by the attempt to increase the money supply and are sufficient to determine the price level, the interest rate, and real wealth.

[19]Ronald McKinnon and Wallace Oates were among the first to use stationary state models. An example of more recent contributions is Johan Myhrman. Michael Parkin has surveyed the progress made in analyzing the stationary state and provides references to models which allow for growth in the labor force and technical change.

[20]The model possesses this property even if the home central bank sterilizes the money supply completely, for then the stock of reserves, while no longer an argument in the excess demand for money, appears in the excess demand for home securities. It can be shown that the greater the proportion of changes in reserves that is sterilized, the larger the absolute change in reserves following a shock to stationary state equilibrium. Only if home and foreign securites are perfect substitutes and the central bank sterilizes the money supply completely is there no automatic adjustment mechanism. In this extreme case saving, the current account, and the excess demand for money must all equal zero, but there are only two endogenous variables, the price level and real wealth.

[21]The existence of a unique and stable stationary state equilibrium is assumed. The two state variables of the system are real wealth and the money wage. The time derivative of real wealth is real saving. It is assumed that the time derivative of the money wage depends positively on the excess demand for labor which in turn depends negatively on the real wage. Investigation of this system of two differential equations shows that there are sets of parameters for which the stationary state equilibrium is locally stable.

[22]Under flexible exchange rates there is a third state variable, the "long-run equilibrium" exchange rate. The time derivative of this variable is assumed to depend positively on the gap between the actual exchange rate and the current value of the long-run equilibrium exchange rate. Investigation of the system of three differential equations composed of the equation just described and the two described in footnote 21 shows that there are sets of parameters for which the flexible-exchange-rate stationary state is locally stable.

[23]It is assumed that there is no sterilization.

None of the results discussed above are associated with neutrality. Indeed, the illustrative model used so far is not neutral. An economic model may be said to exhibit neutrality with respect to a particular nominal variable if 1) the model is homogeneous of degree zero in all nominal variables, and 2) the nominal variable being changed is the only exogenous nominal variable.[24] The illustrative model can be modified so that it exhibits neutrality for changes in certain variables by assuming that the government issues securities which are claims to one unit of the home good rather than being fixed in nominal value. The model is homogeneous under either assumption, but making government securities real securities removes an exogenous nominal variable.

Particularly simple and intuitively appealing answers to three questions can be obtained from stationary state models like the modified illustrative model which exhibit neutrality. First, under fixed exchange rates the only exogenous nominal variable is the exchange rate, so the model is neutral with respect to a devaluation; in particular, a devaluation has no effect on the trade account in the long run.[25] Second, under flexible exchange rates the only exogenous nominal variable is the money supply, so the model is neutral with respect to an increase in the money supply keeping the real bond supply constant. This second result is the analytical foundation of the relative purchasing power parity doctrine. If an economy in stationary state equilibrium experiences an increase in its money supply not accompanied by any change in behavioral relations or exogenous real variables, it eventually reaches a new stationary state in which the price level and exchange rate are increased in proportion to the increase in the money supply, and the interest rate and real wealth remain unchanged as do the trade account and service account. It is well known that the use of this insight to determine a "long-run equilibrium" exchange rate for use as a focal point for intervention under managed floating or to set a new peg under fixed rates is tricky business for three reasons: 1) the economy may not have been in stationary state equilibrium to start with, 2) important coincident changes of the type assumed away above may have occurred, and 3) the short-run stickiness of some nominal variables such as money wages may imply significant adjustment costs.

Third, policy actions have identical effects on real variables under both fixed and flexible exchange rates. To confirm this result using the illustrative model, note that in the stationary state three equilibrium conditions, the saving condition, the current account condition, and the securites market condition, depend only on the three real variables of the system, the relative price of home goods, the interest rate, and real wealth, and that a given policy action affects the same equilibrium conditions under both exchange rate regimes.[26] The money market equilibrium condition determines recursively the stock of reserves or the exchange rate depending on the exchange rate regime.[27]

IV. Some Unfinished Business

While much is known about the comparative statics properties of open economy portfolio balance models referring to each of the three time horizons considered above, relatively little has been learned about the properties of dynamic portfolio balance models.[28] This situa-

[24]This way of describing sufficient conditions for neutrality is due to Don Roper.

[25]However, if there is any sterilization by the cental bank, the devaluation is no longer neutral since the supply of real government securities is increased in response to changes in reserves

[26]Since under either exchange regime an increase in the supply of money with no change in the real supply of bonds has no effects on the real variables of the system, the effects on these variables of an open market purchase and of a decrease in the real supply of securities with no change in the supply of money are identical

[27]If there is any sterilization by the central bank, the fixed exchange rate model no longer exhibits neutrality; the effects of policy actions are no longer the same under the two exchange rate regimes; and the money market equilibrium condition no longer determines recursively the stock of reserves.

[28]Branson, Dornbusch (1975) and (forthcoming), Kouri, and Jacob Frenkel and Carlos Rodriguez have developed dynamic portfolio balance models

tion is unfortunate because most of the problems which concern us do not exist in the trouble-free stationary state and because the solutions which we implement have consequences which extend beyond the short run. It is important to develop dynamic models in order to study, for example, the transmission mechanisms for monetary and fiscal policy, the generation and propagation of inflation, the process of balance of payments adjustment under alternative sterilization policies, the effects of different methods of financing government budget deficits, the implications of alternative hypotheses regarding expectations formation, and the consequences of various intervention strategies. Because of the inherent difficulty of working with dynamic models that contain more than a few endogenous variables, it is to be expected that dynamic problems will be addressed using a number of special purpose models rather than a single general model.[29]

The task of evaluating policies in a dynamic setting is difficult because there are no generally accepted representations of policy makers' preferences. Some insight can be gained by examining the conditions for local stability of stationary state equilibrium. It would be useful to know that a policy 1) insures that stationary state equilibrium will be reached when stationary state equilibrium is unstable in the absence of the policy or 2) speeds convergence to stationary state equilibrium. The second kind of information would be particularly useful in cases when the variables of interest adjust monotonically to their stationary state values. However, while the legacy of a policy for later time periods is important, it is what happens in the first few quarters or years after a policy is undertaken that is of greatest interest, so a study of the stability properties of stationary state equilibria will not suffice.

[29]Parkin has emphasized this point.

REFERENCES

Stanley Black, *International Money Markets and Flexible Exchange Rates*, Studies in International Finance No. 32, Princeton University 1973.

Russel Boyer, "Commodity Markets and Bond Markets in a Small, Fixed-Exchange Rate Economy," *Can. J. Econ.*, Feb. 1975, *8*, 1–23.

William Branson, "Stocks and Flows in International Monetary Analysis," in A. Ando, R. Herring, and R. Marston, eds., *International Aspects of Stabilization Policy*, Federal Reserve Bank of Boston Conference Series No. 12, Boston 1974.

Rudiger Dornbusch, "Exchange Rate Dynamics," *J. Polit. Econ.*, forthcoming.

———, "A Portfolio Balance Model of the Open Economy," *J. Mon. Econ.*, Jan. 1975, *1*, 3–20.

J. Marcus Fleming, "Domestic Financial Policies Under Fixed and Floating Exchange Rates," *I.M.F Staff Papers*, Nov 1962, *9*, 368–79.

Charles Freedman, "A Model of the Eurodollar Market," *J. Mon. Econ.*, forthcoming.

Jacob Frenkel and Carlos Rodriguez, "Portfolio Equilibrium and the Balance of Payments. A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674–88.

Lance Girton and Dale Henderson, "Central Bank Operations in Foreign and Domestic Assets Under Fixed and Flexible Exchange Rates," in P. Clark, D. Logue, and R. Sweeney, eds., *The Effects of Exchange Rage Changes*, forthcoming.

John Hewson and Eisuke Sakakibara, *The Eurocurrency Markets and Their Implications*, Lexington 1975.

Harry Johnson, "The Monetary Approach to Balance of Payments Theory," in *Further Essays in Monetary Economics*, London 1972, 229–49.

Pentti Kouri, "The Exchange Rate and the Balance of Payments in the Short Run and Long Run: A Monetary Approach," *Scand. J. Econ.*, 1976, *78*, 280–304.

Ronald McKinnon and Wallace Oates, *The Implications of International Economic Intergration for Monetary, Fiscal, and Exchange Rate Policy,* Studies in International Finance No. 16, Princeton University 1966.

Robert Mundell, *International Economics,* New York 1968.

Michael Mussa, "A Monetary Approach to Balance of Payments Analysis," *J. Money Credit and Banking,* Aug. 1974, *6,* 333–51.

Johan Myhrman, *Monetary Policy in Open Economies,* Monograph Series, No. 5, Institute for International Economic Studies, Stockholm 1975.

Michael Parkin, "Macro-Models of the Open Economy: A Critical Survey," Discussion Paper, University of Western Ontario, London, Ontario, March 1976, processed.

Don Roper, "Two Ingredients of Monetarism in an International Setting," Seminar Paper 46, Institute for International Economic Studies, Stockholm, April 1975, processed.

Neil Wallace, "The Determination of the Stock of Reserves and the Balance of Payments in a Neo-Keynesian Model," *J. Money, Credit and Banking,* Aug. 1970, *2,* 269–90.

RECENT CONTROVERSIES IN MONETARY THEORY

Irving Fisher and Autoregressive Expectations

By JOHN RUTLEDGE*

The purpose of this paper is to show that Irving Fisher's work on inflation expectations—from *Appreciation and Interest* (1896) through *The Theory of Interest* (1930)—has been seriously misinterpreted in the recent literature. This misinterpretation will be attributed to the failure of recent writers to recognize Fisher's key distinction between situations in which inflation is fully anticipated (full equilibrium) and situations in which inflation is not fully anticipated (the transition period). Fisher's empirical work was an attempt to examine the disequilibrium properties of the interest-inflation relationship. Modern writers, in contrast, interpret Fisher's empirical work as evidence concerning his theory of appreciation and interest in full equilibrium.

In particular, two propositions central to the modern interpretation of Fisher's interest theory were not, in fact, held by Fisher. Fisher did NOT assume that real rates of interest are determined independently of past inflation rates, and did NOT interpret the distributed lag weights in regressions of nominal interest on past inflation as autoregressive price expectations parameters.

I. Fisher's Full Equilibrium Theory

Fisher stressed that the effect of appreciation on interest rates depended fundamentally on whether or not the appreciation was foreseen, and chose to present the theoretical relation in a model assuming full loan market equilibrium, i.e., perfect foresight.

> We must begin by noting the distinction between a foreseen and unforeseen change in the value of money . . At present we wish to discuss what will happen, ASSUMING THIS FORESIGHT TO EXIST [1896, p. 6]

> The rate of interest in the relatively depreciating standard is equal to the sum of three terms, viz., the rate of interest in the appreciating standard, the rate of appreciation itself, and the product of these two elements,
>
> $$j = i + a + ia$$
>
> Thus to offset appreciation, the rate of interest must be lowered by slightly more than the rate of appreciation [1896, p. 9]

The foregoing analysis, that interest rates will rise by the expected rate of inflation, is the essence of most modern interpretations of Fisher's work. Indeed, the proposition that real interest rates remain constant while the nominal rate adjusts for expected inflation is known as the "Fisher Relation" Modern writers, unfortunately, typically fail to note the assumption of perfect foresight, and apply this analysis to disequilibrium situations. As we shall see in Section III, Fisher supplied a very thorough analysis of the effects of imperfectly foreseen inflation on interest rates during the transition period

Empirical evidence presented by Fisher to test the theory of interest adjustment under perfect foresight yielded two important results. Market interest rates were high during periods of inflation and low during periods of deflation. Thus, on qualitative grounds, the theory must be accepted. Fisher found, however, that interest rates responded slowly and incompletely to inflation. The rate of interest expressed in commodities was many times more variable than the

*Associate Director, Applied Financial Economics Center, Claremont Men's College; and International Economist, Treasury Department

money rate and was often negative during periods of rapid inflation. Since traders had the option of simply hoarding commodities—and earning zero commodity interest—Fisher concluded that the evidence "must mean that price movements were inadequately foreseen" (1896, p. 67). Fisher analyzed the inadequacy of interest adjustment largely in terms of money illusion.

> When prices begin to rise, money interest is scarcely affected. It requires the cumulative effect of a long rise, or of a marked rise in prices, to produce a definite advance in the interest rate. If there were no money illusion and if adjustments of interest rates were perfect, unhindered by any failure to foresee future changes in the purchasing power of money or by custom or law or any other impediment, we should have found a very different set of facts [1896, p 416]

II. Fisher's Theory of the Transition Period

Forced to reject the perfect foresight model, Fisher presented an alternative model of the effects of inflation on nominal interest rates based on imperfect foresight, i.e., a disequilibrium model of interest adjustment in which transitory changes in real variables—profits, investment, volume of trade—play a central role.

Fisher's view of interest adjustment in *Appreciation and Interest* is based on his empirical judgments that firms are net debtors, that owners of firms are endowed with foresight superior to that of bondholders, and that entrepeneurs forecast profits autoregressively.

> Suppose an upward movement of prices begins Business profits (measured by money) will rise, for profits are the difference between gross income and expense, and if both these rise, their difference will also rise. Borrowers can now afford to pay higher "money interest " If, however, only a few persons see this, the interest will not be fully adjusted and borrowers will realize an extra margin of profit after deducting interest charges This raises an expectation of a similar profit in the future, and this expectation, acting on the demand for loans will raise the rate of interest [1896, p 75]

Thus borrowers, expecting to pay back loans in depreciated currency, will increase their demand for loans. Lenders, however, are unable to forecast the inflation; hence—as long as the supply of loans is not perfectly inelastic—the rate of interest will only be partially adjusted upwards. The result is that firms pay lower actual commodity interest on their borrowings, and there is a windfall increase in profits. This leads firms to revise forecasts of future profits upward, hence increases the demand for loans still further.

In summary, Fisher argued that the transition period would be characterized by an increase in the nominal rate, a decrease in the realized commodity rate of interest, an increase in real business profits and an increase in the real value of investment. Fisher regarded this abnormally low commodity rate of interest and the resulting overinvestment as the major determinant of the "trade cycle."

> In the view here presented periods of speculation and depression are the result of *inequality* of foresight If all persons underestimated a rise of price in the same degree the non-adjustment of interest would merely produce a transfer of wealth . It would not influence the volume of loans . . In the actual world, however, foresight is very unequally distributed. Only a few persons have the faculty of always "coming out where they look." Now it is precisely these persons who make up the borrowing class. Just because of their superior foresight, society delegates to them the management of capital. It is they who become "captains of industry." . . . . It therefore happens that when prices are rising, borrowers are more apt to see it than lenders. . . This will of course raise the rate of interest But it will also cause an increase of loans and investments This constitutes part of the stimulation to business which bimetallists so much admire. [1896, p. 77]

We see, then, that Fisher's theory of the transition period is of a qualitatively different sort than the theory of perfect foresight described in Section I. Fisher recognized that during the adjustment period changes in the rate of inflation will have important effects on *real* variables, commodity interest, the rate of investment spending, and the volume of trade, rather than causing a simple adjustment of the money rate.

In the *Purchasing Power of Money* Fisher presented his most thorough analysis of the dynamics of interest adjustment.

> As prices rise, profits of business men, measured in money will rise also, even if the costs of business were to rise in the same proportion . . . But, as a matter of fact, the business man's profits will rise more than this because the rate of interest he has to pay will not adjust itself immediately. Among his costs is interest, and this cost will not, at first, rise. Thus the profits will rise faster than the prices. Consequently he will find himself making greater profits than usual, and be encouraged to expand his business by increasing his borrowings. [1912, pp. 58–59]

Since Fisher's primary concern here was the relationship between inflation and the trade cycle, he paid particular attention to the effects of interest adjustment on real investment and the volume of trade. The sequence of interest adjustment, credit expansion, and inflation is formally similar to Wicksell's cumulative process, which Fisher appropriately cited. Fisher argues:

> These borrowings are mostly in the form of short-term loans from banks, and, as we have seen, short-term loans engender deposits. As is well-known, the correspondence between loans and deposits is remarkably exact. Therefore deposit currency (M′) will increase, but this extension of deposit currency tends further to raise the general level of prices. . Hence prices, which were already outstripping the rate of interest, tend to outstrip it still further, enabling borrowers, who were already increasing their profits, to increase them still further. More loans are demanded, and though nominal interest may be forced up somewhat, still it keeps lagging below the normal level. Yet nominally the rate of interest has increased; and hence the lenders, too, including banks, are led to become more enterprising. Beguiled by the higher nominal rates into the belief that fairly high interest is being realized, they extend their loans, deposit currency (M′), already expanded, expands still more . . . . In other words, a slight initial rise of prices sets in motion a train of events which tends to repeat itself. Rise of prices generates rise of prices, and continues to do so as long *as the interest rate lags behind its normal figure*. [1912, pp. 59–60]

These effects, together with the decrease in the demand for money due to higher nominal interest rates (1912, p. 63), result in a temporary increase in real output, due to the excess of investment over savings. Thus the interest rate is unable to adjust to the normal level and Fisher concludes that,

> Trade . . . . will be stimulated by the easy terms of loans. This effect is always observed during rising prices, and people note approvingly that "business is good" and "times are booming." [1912, p. 61]

Unfortunately, once the adjustment of interest to its normal level is completed, the stimulus to trade is removed. In fact, Fisher maintains, the lags in interest adjustment may precipitate a contraction of trade.

III. Fisher's Empirical Work

In each of his books on interest rates Fisher devotes an early chapter to a discussion of the steady-state properties of the interest inflation relationship. Fisher felt that steady-state—or full equilibrium—properties were of great interest. Long-run conditions provide us with information about where a given system *would* come to rest, if all exogenous variables were to remain fixed; hence, they help us to identify the major directions of adjustment in endogenous variables to expect, which may ultimately aid us in formulating policies.

Fisher did not, however, call on the steady-state properties of interest rates to explain any given set of observed market interest rates. The world is simply not kind enough to change slowly enough to observe the steady state.

> . . . we shall consider the temporary effects during the period of transition separately from the permanent or ultimate effects These permanent or ultimate effects follow after a new equilibrium is established,—if, indeed, such a condition as equilibrium may be said ever to be established [1930, pp 55–56]

Thus, Fisher stressed, the *observable* world is in continual disequilibrium, and the appropriate framework for analyzing real-world data is dynamic—or transition period—rather than static.

> . . . practically there is almost always some occurrence to prevent perfect equilibrium Oscillations are set up which, though tending to be self-corrective, are continually perpetuated by fresh disturbances. . . . A ship in a calm sea will 'pitch' only a few times before coming to rest, but in a high sea the pitching never ceases. While continually seeking equilibrium the ship

> continually encounters causes which accentuate the oscillation . . . Since periods of transition are the rule and those of equilibrium the exception, the mechanism of exchange is almost always in a dynamic rather than a static condition. [1930, pp. 70–71].

Fisher accordingly took great care to separate his discussions of steady-state properties from those of empirical work on interest adjustment.

Fisher's most sophisticated, and most frequently cited, empirical work on the effects of inflation on interest was presented in Chapter XIX of *The Theory of Interest*. In the first five sections of this chapter Fisher gave qualitative evidence showing that interest rates expressed in different standards are different when those standards are diverging in value; inflation resulted in high interest rates, but the adjustment was very slow. Then, in Sections 6 and 7, Fisher presented results of correlating nominal interest rates with lagged inflation rates for both Great Britain and the Unites States. He convincingly argues (pp. 418–20) that a distributed lag of past inflation rates should be employed rather than a discrete lag, and presented simple correlation coefficients which would correspond to a regression equation of the form,

$$(1) \qquad r_t = a + b \sum_{i=1}^{T} w_i \pi_{t-i} + e_t .$$

where a and b are constant, $w_i, i = 1, \ldots, T$ are distributed lag weights, T is the order of the estimated lag distribution, and e_t is the error term. For ease in calculation Fisher constrained the lag weights to decline arithmetically and to sum to unity. The results were striking Correlations between long-term bond rates and distributed lags of past changes were extremely high. Even more striking, however, was the length of time required for complete adjustment. After estimating (1) for various values of T, Fisher concluded that,

> for Great Britain in 1898-1924, the highest value of r (+0.980) is reached when effects of price changes are assumed to be spread over 28 years or for a weighted average of 9.3 years, while for the United States the highest r (+0.857) is for a distribution for the influence due to price changes over 20 years or a weighted average of 7.3 years. [1930, p. 423]

Using quarterly data on *U.S.* commercial paper rates Fisher found similar results:

> . . . . in the period 1915-27, r reaches its maximum (+0.738) only when a total of 120 quarters, or thirty years, is included in the period subject to the influence of price changes upon i. [1930, p. 427]

In contrast to modern writers, Fisher was not surprised to find that it took, in the most extreme case, 120 quarters for full adjustment of interest to inflation. He interpreted the lag as largely representing adjustments in real variables, real interest, profits, and the volume of trade, and *not* as a simple measure of the delay in expectations formation.

> It seems fantastic, at first glance, to ascribe to events which occurred last century any influence affecting the rate of interest today And yet that is what the correlations with the distributed effects of (inflation) show. A little thought should convince the reader that the effects of bumper wheat crops, revolutionary discoveries and inventions, Japanese earthquakes, Mississippi floods, and similar events project their influence upon prices and interest rates over many future years *even after the original causal event has been forgotten* . . . A further probable explanation of the surprising length of time by which the rate of interest lags behind price changes is that between price changes and interest rates a third factor intervenes. This is business, as exemplified or measured by the volume of trade. It is influenced by price changes and influences in turn the rate of interest. [1930, p. 428–29]

On this point Fisher further elaborates,

> Two facts have, I think, now been well established The first, that price changes influence the volume of trade, has been shown in earlier studies made by me The second, that the volume of trade influences the rate of interest, has been shown by Carl Snyder, Ayres, Mitchell, and others The evidence for both relationships is not only empirical, but rational Rising prices increase profits both actual and prospective, and so the profit taker expands his business His expanding or rising income stream requires financing and increases the demand for loans [1930, p. 439]

Fisher sums up the adjustment of both real and nominal interest.

> The final result, partly due to foresight and partly to the lack of it, is that price changes do after several years, and with the intermediation of changes in profits and business activity affect interest very profoundly. In fact, while the main purpose of this book has been to show how the rate of interest would behave if the purchasing power of money were stable, *there has never been any long period of time during which this condition has been approximately fulfilled When it is not fulfilled, the money rate of interest, and still more the real rate of interest, is more affected by the instability of money than by those more fundamental and more normal causes connected with income, impatience, and opportunity, to which this book is chiefly devoted.* [1930, p. 451]

Quite clearly, then, Fisher did not assume a constant real rate, nor did he assume a real rate determined independently of past inflation. It is precisely the variations in real factors, according to Fisher's interpretation, which combine to produce the extremely long adjustment period for nominal interest, quite apart from the way in which price expectations are formed

During the past decade a large body of empirical work on the interest inflation relationship has been published These papers typically presented estimates similar to (1) in form, and were widely interpreted as a revival of the work begun by Fisher in 1930 The interpretation of the empirical work, however, was much more naive than that given by Fisher. In fact, in their zeal to obtain estimates of price expectations parameters many writers went so far as to argue that Fisher assumed a real rate which was determined by "productivity and thrift" alone and which was independent of past inflation rates

But *Fisher did not interpret the lag between inflation and interest in terms of price expectations* As I have attempted to show above, Fisher felt the major influence of inflation was on real interest, and that the adjustment came largely through profits, investment, and the volume of trade.

> As to the nominal variation in the rate of interest, we found that theoretically, an appreciation of one percent of the standard of value in which the rate of interest is expressed, compared with some other standard, will reduce the rate of interest in the former standard, compared with the latter, by about one percent, . Such a change in the rate of interest would merely be a change in the number expressing it, and not fundamentally a real change. *Yet, in actual practice, for the very lack of this perfect theoretical adjustment, the appreciation or depreciation of the monetary standard does produce a real effect on the rate of interest, and that a most vicious one This effect, in times of great changes in the purchasing power of money, is by far the greatest of all effects on the rate of interest.* [1930, p 493]

The current interpretation of Fisher's work, however, continues to ignore the effects of inflation on real interest, and to interpret the distributed lag weights in (1) as the parameters of the inflation forecasting mechanism.

IV. Conclusions

The purpose of this paper has been to demonstrate that Fisher's theory of appreciation and interest was much more theoretically innovative and far richer in implications than the version found in the literature today The modern interpretation of Fisher's work is based on two propositions, both of which are false. 1) Fisher assumed a constant real rate, or a real rate which was orthogonal to past inflation rates; 2) Fisher's distributed lag estimates constituted an attempt to measure the parameters of the inflation forecasting process, implying that long lags are associated with slow revision of price forecasts.

Fisher's theory of appreciation and interest, as he advised his readers many times, was based on the crucial distinction between periods of *full equilibrium* and those of *transition*, or disequilibrium. Fisher explained that in steady-state equilibrium, nominal interest will be bid up by exactly the rate of inflation, and real interest will remain unchanged. But the overwhelming impact of inflation during the transition period is on real variables: real interest, real profit, real investment and real income. In fact Fisher argued that the *real effects* of inflation on interest were the major determinants of booms in business activity.

The neglect of Fisher's rigorous development

of dynamic interest theory is unfortunate for several reasons. It means, first of all, that an important chapter has been left out of modern views on the development of economic thought. Fisher is typically cast as a stubborn classical economist who stuck to his classical comparative static principles of full employment equilibrium, and perfectly flexible prices and interest rates. Yet we have seen how Fisher was fundamentally concerned with statistical analysis of real-world data, and how he skillfully developed a dynamic model of interest and income adjustment which was based on imperfect foresight, money illusions, windfall profit affecting aggregate demand and income, and tardy adjustment of prices—all characterizing the transition period to a new static equilibrium. Thus the work of Irving Fisher must be interpreted as a pioneer effort at understanding the short-run dynamics of macroeconomic disequilibrium, an interpretation recently given to Keynes' *The General Theory* by Axel Leijonhufvud, and others. Indeed, once Fisher's work is recognized, *The General Theory* can no longer be viewed as a hiatus in the development of modern economics, but rather as a continuation of work begun by Fisher several years earlier

Of more practical importance is the fact that an entire body of literature has grown out of a false interpretation of Fisher's empirical work in *The Theory of Interest*, a body of literature which unanimously interprets long interest-inflation lags in terms of sluggish forecast formation, perhaps resulting in incorrect policy prescriptions. A further consequence of the modern interpretation of Fisher's work, has been the proliferation of autoregressive expectations proxies for a variety of purposes, rather than examining alternative forecasting models.

REFERENCES

John B. Clark, "The Gold Standard in the Light of Recent Theory," *Political Science Quarterly*, Sept. 1895.

William Douglass, "A Discourse Concerning the Currencies of the British Plantations in America," Boston 1740.

Pierre des Essars, *Journal de la Societe de Statistique de Paris*, April 1895.

Irving Fisher, *Appreciation and Interest*, New York 1896

———, *The Rate of Interest*, New York 1907.

———, "The Business Cycle Largely A 'Dance of the Dollar'," *J. Amer. Statist. Assn.*, Dec. 1923, 1024–28.

———, "Our Unstable Dollar and the So-Called Business Cycle," *J. Amer. Statist. Assn.*, June 1924, 179–202

———, *The Purchasing Power of Money*, New York 1926.

———, *The Theory of Interest*, New York 1930.

Jacob de Haas, "A Third Element in the Rate of Interest," *Journal of the Royal Statistical Society*, March 1889.

John Stuart Mill, *Principles of Political Economy*, (ed.) Ashley, London 1923.

J. Pease Norton, *Statistical Studies in the New York Market*, New York 1902.

Knut Wicksell, "Der Bankzins als Regulator der Warenpreise," *Jahrbucher fur Nationalokonomie*, 1897, *68*, 228–43.

The Anatomy of Monetary Theory

By ROBERT W. CLOWER*

It is obvious even to the most ordinary intelligence, that a commodity should be given up by its owner in exchange for another more useful to him. But that every economic unit in a nation should be ready to exchange his goods for little metal disks apparently useless as such, or for documents representing the latter, is a procedure so opposed to the ordinary course of things [as to seem] downright "mysterious."

—Karl Menger, "On the Origin of Money"[1]

Modern discussions of monetary theory have fairly well demolished its traditional foundations without so far putting anything definite in their place. To be sure, some progress towards reconstruction has been made. In particular, contributions by John R. Hicks (1967), F. H. Hahn (1971), Karl Brunner and Allen Meltzer, Jack Hirschleifer and a host of other writers have shown that set-up costs of engaging in trade must play a central role in any acceptable formal theory of monetary exchange. Other contributions by Clower (1969), Morris Perlman, Joel Fried, Joseph Ostroy and R. M. Starr, and Peter Howitt have suggested that individual holdings of commodity inventories must also be taken into account. Finally, many writers have argued that certain physical characteristics of commodities are vital. Thus some of the ingredients for a reconstruction of the foundations are not seriously in doubt. What remains to be settled is the manner in which these and possibly other ingredients might best be combined.

This problem is not likely to be resolved in the near future; at best, we might hope for an early professional consensus about the way in which the problem should be approached. My purpose in this paper is to contribute to the formation of such a consensus by exploring rather carefully the roles that transactions costs and other purportedly crucial complications might play in the evolution of monetary exchange arrangements within a simple spot-exchange economy that is initially imagined to be as devoid of explicit empirical features as any standard Arrow-Debreu model.

I

Our first order of business is to state explicitly just what aspects of experience we should want to have "explained" by any theory that claims to provide even a minimally adequate description of a monetary economy. This may be accomplished most conveniently by setting out an agenda of requirements that any such theory should satisfy. On the basis of earlier literature, everyday observation and my own considered professional judgment, I should regard four requirements as mandatory:

1) The theory should imply that trade is an ongoing process in time rather than a once-for-all affair that ends in the permanent elimination of incentives for further trade.

2) The theory should imply that, on average over any finite time interval, each individual holds positive stocks of all goods that are regularly traded.

3) The theory should imply that the bulk of all trades occur not through essentially ran-

*University of California-Los Angeles. I am indebted to numerous colleagues for helpful personal discussions about central ideas in this paper, but especially to Joel Fried and Peter Howitt of the University of Western Ontario, and to Robert Jones and John Riley of UCLA. As will become clear in the sequel, all of my thinking in this area has been strongly influenced by recent work of Professor Hicks, particularly his *Theory of Economic History* (1969) and *Critical Essays in Monetary Theory* (1967). My debt to other writers, both at UCLA and elsewhere, is not too inadequately acknowledged, I hope, by citations in the text.

[1]Menger, p. 239. I owe this reference to Robert Jones, whose Brown University dissertation supplies one possible solution to Menger's "paradox" (for a published version of the argument, see, 1976).

[2]For extensive discussion and references see A. M. Ulph and D. T. Ulph.

dom pairing of individuals who happen to share a double coincidence of wants, but rather through systematic pairing of specialist with nonspecialist traders in a relatively small number of organized, continuously operating markets.

4) The theory should imply that at least one and at most a few distinctive "money" commodities are transferred (or promised for future delivery) by one party to another in virtually all exchange transactions.

The rationale of each of these conditions will become clear as we proceed.

II

Directing inquiry now to our central theme—the logical anatomy of monetary theory—let us start by imagining a Patinkinesque community of self-interested individuals each of whom receives "like manna from heaven" a predetermined quantity per time unit of one or more durable goods that may be consumed directly, traded for other commodities, or held for future consumption or trade. Suppose also that each individual is a natural source of labor services that can be consumed directly (as "leisure" or as inputs in household "production") or contracted for sale to other individuals. Given these assumptions, received theory informs us that potential gains from trade will exist if different individuals have different preferences or endowments. Let us assume that this proviso is satisfied. Then, depending on the magnitude and distribution of potential gains, trades will almost certainly occur. What we can validly say about the volume and pattern of trading activity will then depend on just what story we choose to tell about the manner in which individual trading plans are conceived and executed.[3]

At this stage, of course, we have almost unlimited room for maneuver. Our basic model includes certain necessary elements of an acceptable theory of monetary exchange, namely, objects to be traded, agents to trade these objects, and incentives for agents to interact so that actual trades occur. But it lacks all elements that might be or have been adduced by various writers as necessary conditions for monetary exchange. Our next task, therefore, is to add further content to our model.

III

Let us proceed by asking, first, what additional assumptions should be included in our argument to ensure satisfaction of the first of the requirements listed earlier, i.e., the requirement that trading activity should occur as an ongoing process in calendar time?

At first glance, the answer to this question might seem to be, "none at all," since our basic model is explicitly formulated as a stock-flow system in which potential gains from trade are sustained over time by the continuous receipt of fresh commodity endowments by individual traders. As Roy Radner, Hahn and others have shown,[4] however, even outwardly "essential" sequence models may turn out to be logically equivalent to "nonessential" sequence models of Arrow-Debreu type in which trading contracts are concluded at just one instant in calendar time. The source of this equivalence is significant; it lies in the twin assumptions that traders are inhumanly prescient and that trading contracts and arrangements for future delivery of commodities can be negotiated at zero cost. If either or both of these restrictions are dropped, the first of our requirements can be easily satisfied by almost any stock-flow model.[5]

Let us drop both restrictions; more precisely, let us suppose that individuals view future endowment flows as probable rather than certain, and also that individuals can negotiate trades

[3]This theme is most elegantly elaborated in papers by E. C. H. Veendorp, Ostroy (1973), Allan Feldman, P. J. Madden and Jones.

[4]Cf. Radner, Hahn, pp. 230–34, Ulph and Ulph, pp. 365–67.

[5]Strictly speaking, pure stock models fail this test, but such models are seldom used except to discuss special aspects of the logistics of exchange, for which purpose they are invaluable (cf. Ostroy).

only by engaging in extensive search and bargaining activities. Then our model will satisfy our first requirement. Will it also satisfy the second; i.e., can we assert that each individual will hold positive average stocks of all goods that are regularly traded?

Again, the answer would appear at first thought to be in the affirmative. For it is easy to show that if search, bargaining and other trade-related activities impose set-up costs on individuals in the form of foregone leisure or consumption or both, then individuals will engage in trade, if at all, only at discrete points rather than continuously in time; hence, each individual will hold, at one moment of time or another, positive stocks of all traded goods.[6] What we have to show, however, is not that holdings of stocks will occasionally be positive, but rather that average holdings over any finite time interval will be positive.

To avoid tacitly assuming what we seek to prove, let us suppose that we have to deal with an individual who wants to trade a good X for another good Y, but discovers to his horror that no other individual is willing to trade X or Y for anything but a third good, say, Z. In this case the individual will either not trade at all, or he will first sell X for Z and then use Z to purchase Y. Unless units of Z are less costly to store than units of either X or Y, however, the individual will (if rational) combine every sale of X or Z with a nearly simultaneous purchase of Y with Z, which implies that average holdings of units of Z will be arbitrarily close to zero. Thus storage costs as well as set-up costs apparently must be invoked to ensure that our model satisfies the second requirement listed earlier. Actually, this is too strong. If search and bargaining costs are substantial relative to potential gains from trade, then indirect trades will be infrequent or nonexistent, in which case storage costs will be irrelevant. But there is also another possibility. If individuals regard future endowment flows as merely probable rather than certain, we should expect many or even most traders to hold precautionary balances of some goods as a hedge against possible real-income reductions, and such balances would consist predominantly of goods with relatively low costs of storage. If such holdings are widespread and are concentrated in a few goods that are held by virtually all individuals, then it may well cost less time and effort for an individual to trade X for one of these goods, and then go on to purchase Y, than to trade X for Y directly. In this situation, therefore, our argument goes through as stated initially: i.e., storage as well as transactions costs must be invoked to ensure satisfaction of our second requirement.

IV

The third requirement in our agenda—that the bulk of all trades should occur in organized markets—appears not to have been noticed in earlier work or, as seems more likely, has simply been taken for granted.[7] However that may be, the oversight is crucial; for as we shall see later, the existence of organized markets appears to be almost a precondition for monetary exchange. But our immediate task is just to explain how, on our present assumptions, organized markets might evolve in response to the working of natural economic forces.

If potential gains from trade were large and widely distributed across individuals and commodities, trading activity might be substantial even in a community without organized markets; but if search and bargaining costs were at all significant, vast areas of potential gain would never be explored. In the latter case, as George Stigler has observed,[8] there would exist "powerful inducements"—namely, widespread profit opportunities—for some individuals to localize trade by establishing "ready markets"

[6] See Hirshleifer, pp. 138-41, for an especially clear exposition.

[7] This is reflected in the common use of the word "money" to refer to a complex of ideas that would be more accurately rendered by using the term "organized markets." For more on this, see below, note 11.

[8] Stigler, pp. 218-19.

for commonly traded goods in which other individuals could routinely execute certain designated pairwise trades at dates and in quantities of their own choosing.

Powerful inducements notwithstanding, we should not expect many individuals to act as specialist traders. To establish a ready market and attract enough customers to earn a profit that would make the activity worthwhile, an individual would have to:

(i) accumulate inventories of a wide variety of commonly traded goods;
(ii) offer to trade at rates of exchange more attractive than average rates which individuals could normally expect to obtain on short notice by trading with other nonspecialists;
(iii) maintain a spread between "buy" and "sell" rates that would encourage volume trading and discourage competition by other specialists, yet yield a real income sufficient to offset operating costs;
(iv) earn a rate of return on average holdings of trade inventories at least equal to his rate of time discount.

Casual introspection suggests that these conditions would dissuade any but the most thrifty, foresighted, diligent, energetic, sagacious and enterprising individuals[9] from setting up shop as trade specialists. But specialist traders could never be common in any case; for if they tended in that direction, competition among them would reduce trading spreads to the point where only those specialists with relatively low rates of time discount would choose to survive.

I need not emphasize the social benefits that flow from the existence of organized markets. The trouble and effort that would otherwise be incurred in the conduct of the most ordinary business of life has been a favorite theme of writers on money since the time of Aristotle.[10] What does merit emphasis is that these benefits do not depend logically on the use of special "money" commodities as media of exchange. The literature of monetary economics—again from Aristotle on—is replete with instances in which writers have inadvertently used the word "money" as if it were synonymous with the phrase "organized markets."[11] Historically, of course, "money" and "markets" have generally coexisted, but that connection is one of fact rather than logic. Confused understanding of this point is, I conjecture, a major reason why monetary theory has for so long remained one of the least settled branches of formal economic analysis. As we have just seen, it is easy to explain how organized markets arise from the working of natural economic forces; there is no mystery here, except perhaps Adam Smith's "instinct to truck and barter." What is not easy to explain is how the organization of such markets tends always to take a highly specialized form that permits us objectively to assert that certain objects (or "documents representing the latter") play a distinctive role as "money." That, of course, is the task set for us by the fourth and final item on our agenda.

V

Given the individualist behavior assumptions of standard theory, we have no option but to suppose that monetary exchange will emerge, if at all, only if that way of organizing trading activity is clearly advantageous either to specialist traders or to nonspecialist traders and is disadvantageous to neither.

Can it be shown that a typical specialist will have reason to require that customers give or re-

[9]This list is lifted, of course, from various places in Marshall's *Principles*

[10]See A E Monroe, p 17, for Aristotle's contribution, and J H McCulloch, pp 1-2, for a particularly lucid modern version of the same story

[11]This usually occurs as soon as a writer has (with little difficulty) pointed out the costs of simple barter and passed on to consider other cases, for the only other cases that ever seem to come to mind are those that involve fully monetized market exchange For an explicit account of other conceivable cases, see Clower and Axel Leijonhufvud, pp 184–85.

ceive units of a few designated "money" commodities in exchange for all other goods? The answer appears to be in the negative; for although it is known that economies of scale may be achieved by concentrating inventories in relatively high-volume lines of activity,[12] it can be shown that the elimination of one commodity from a specialist trader's list of "tradeable goods" will necessarily reduce trading volume in one or more other commodities that remain on the list—which works in the wrong direction. Indeed, granted that inventory holdings are a potential source of scale economies, and ignoring such considerations as costs of travel and transport and diseconomies in the operation of large-scale markets, the logic of our analysis leads us to conclude that the only permanently viable form of market organization would be one in which a single specialist trader—presumably one with a very low rate of time discount and a very keen eye for profit—provided the only ready market in the entire community. This form of organization would also be socially optimal in the usual efficiency sense since, with all trading activity concentrated in a single market, potential economies of scale in inventory holdings could be exploited to the fullest possible extent.

This line of argument does not demonstrate that a monetary form of market organization would be disadvantageous to specialist traders. However, it strongly suggests that no specialist trader would find it worthwhile to initiate moves in that direction. What about nonspecialist traders? Have they any incentive voluntarily to restrict their dealings to a relatively small subset of the set of all pairwise trades that specialist traders in the aggregate would be able and willing to handle?

Again, the answer appears to be in the negative. For why should any individual who wishes to trade, say, X for Y, proceed instead first to trade X for another commodity Z, and then to trade Z for Y, particularly if it is always possible to trade X directly for Y with less time and effort? On closer inspection, however, this argument is seen to beg the question, for it rests on the tacit assumption that an individual who wishes to trade one good for another also wants to acquire the second good *immediately*. To appreciate why this might not always (or even usually) be so, suppose that an individual's only endowment flow consists of units of a good X, and that all units of this good are sooner or later used to finance purchases of a variety of other goods—$Y,Z,A,B, \ldots$, etc. Then, as has been shown in recent work on the demand for trade inventories and the timing of exchange transactions,[13] the individual will—except in very special cases—minimize total trade-related costs by purchasing many or even most consumption goods at dates that differ significantly from those at which he sells units of X. To be sure, units of at least one consumption good must necessarily be acquired every time units of X are sold. But suppose that X is typically traded in large-size lots in exchange for just one other good—any good will do. Then units of this other good can later be used to purchase yet other commodities in various lot sizes and at various dates to conform with cost and other considerations that underlie the individual's choice of transaction dates for different goods. So we conclude that individuals will quite generally choose to carry out two transactions to go from X to Y, even though only one transaction is ever strictly necessary.[14]

The last result resolves our central problem; for now we have only to recall our earlier discussion of conditions in which certain commod-

[12] See Kenneth Arrow, S. Karlin and Herbert Scarf, pp 7–8

[13] See Herschel Grossman and Andrew Policano for discussion and references, also Clower and Howitt

[14] An obvious exception would occur if "time taken to go to market" greatly outweighed all other trade-related costs, since in this case an individual would "bunch" sales of endowment goods with purchases of goods for consumption, thereby minimizing the number of trips to market. History provides a possible example of this in dealings of the general merchant of pioneer times with trapper and farmer customers. See also Lars Jonung, for some fascinating comments on Swedish experience.

ities would be held and used for transactions purposes to arrive at an obvious and compelling reason why nonspecialist traders might voluntarily restrict themselves to selected pairwise trades, namely, storage costs. If there exists some commodity that is already a common object of exchange and which has distinctly lower storage costs than all or most other commodities, a rational individual will choose to acquire or dispose of units of this good in virtually all exchange transactions. Moreover, since different individuals are unlikely to differ much in their perception of the relative costs of storing different commodities, it follows that nonspecialist traders as a group will choose to conclude most transactions with one or a few goods which will thus come to play a distinctive role as "common media of exchange" and "temporary abodes of purchasing power."[15] If one of these goods is not just inexpensive to store but is also easy to identify, handle, partition, count, hide and transport (Did someone shout "Gold"?), then that good will almost surely dominate all others as a means of payment in spot transactions. But in that case, and perhaps even under weaker conditions, specialist traders will have no incentive to maintain direct pairwise trading of every variety of commodity. Ready markets that deal exclusively with selected commodity lines (groceries, hardware, clothing, meats, black puddings, etc.) will be at least as viable as more general markets, among other reasons because inventory economies of scale can be exploited fully in such markets with relatively small stocks of trade capital. But there is no need to carry the story further; we have already established a satisfactory rationale for the existence and ubiquity of monetary exchange.

[15]If costs of search and bargaining were relatively unimportant, organized markets would not be viable, but the preceding argument would still provide a rationale for individuals to hold and use certain kinds of goods as temporary abodes of purchasing power. Such behavior would probably not be common, however, because the insignificance of transactions costs would encourage frequent sales of small-size lots of endowment goods in direct exchange for goods to be consumed, which would make storage costs a minor factor in the choice of trading dates. An external observer of such an economy would be unlikely, therefore, to see any pattern in the pairing of traded commodities. These patterns become blindingly obvious only in economies where most trades occur in organized markets.

VI

My conclusion can be brief. The preceding analysis indicates that just two main factors, namely, costs of negotiating exchange transactions and certain physical characteristics of commodities, have to be taken into account to establish *necessary* conditions for monetary exchange to emerge in an otherwise strictly Arrow-Debreu economy. *Sufficient* conditions cannot be stated, except in very general terms, because these depend in an essential way on the precise character of individual preferences, on the size and distribution of endowment flows, on the magnitude of search and bargaining costs, and on the technology of market management. Thus our results are not in any sense final or complete; at best they might be said to clarify just how much still remains to be done if we are to make theoretical sense of money.

REFERENCES

Kenneth J. Arrow, S. Karlin and Herbert Scarf, *Studies in the Mathematical Theory of Inventory and Production*, Stanford 1958.

Karl Brunner and Allen Meltzer, "The Uses of Money: Money in the Theory of an Exchange Economy," *Amer. Econ. Rev.*, Dec. 1971, *61*, 784–805.

Robert W. Clower, *Readings in Monetary Theory*, London 1969.

——— **and Peter W. Howitt,** "Money, Credit and the timing of Transactions," UCLA Discussion Paper 72, June 1976.

——— **and Axel Leijonhufvud,** "The Coordination of Economic Activities: A Keynesian Perspective," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 182–88.

A. M. Feldman, "Bilateral Trading Processes, Pairwise Optimality and Pareto Optimality," *Rev. Econ. Stud.*, Oct. 1973, *39*, 463–74.

Joel Fried, "Money, Exchange and Growth," *Economic Inquiry*, Sept. 1973, *11*, 285–301.

Herschel I. Grossman and Andrew J. Policano, "Money Balances, Commodity Inventories, and Inflationary Expectations," *J. Polit. Econ.*, Dec. 1975, *83*, 1,093–1,112

F. H. Hahn, "Foundations of Monetary Theory," in *Essays on Modern Economics*, M. Parkin, ed., London 1973, 230–42

John R. Hicks, "A Suggestion for Simplifying the Theory of Money," *Economica*, Feb 1935, *2*.

———, *A Theory of Economic History*, London 1969.

———, *Critical Essays in Monetary Theory*, Oxford 1967.

Jack Hirshleifer, "Exchange Theory: The Missing Chapter," *Western Econ. J.*, June 1973, *11*, 129–46.

Peter W. Howitt, "Stability and the Quantity Theory," *J. Polit. Econ.*, Jan.-Feb. 1974, *82*, 133–51.

Robert A. Jones, "The Origin and Development of Media of Exchange, *J. Polit. Econ.*, Aug. 1976, *84*, 757–76

Lars Jonung, "The Behavior of Velocity in Sweden, 1871-1913," UCLA Workshop in Monetary Economics Discussion Paper, May 1976.

P. J. Madden, "Efficient Sequences of Non-Monetary Exchange, *Rev. Econ. Stud.*, Oct. 1975, *42*, 581–96.

J. H. McCulloch, *Money and Inflation*, Ne York 1975.

Karl Menger, "On the Origin of Money, *Economic Journal*, June 1892, *2*, 239–55.

A. E. Monroe, *Early Economic Though* Cambridge, Mass. 1927.

J. M. Ostroy, "The Informational Efficienc of Monetary Exchange," *Amer. Econ. Rev.* Sept. 1973, *63*, 597–610.

——— **and R. M. Starr,** "Money and th Decentralization of Exchange," *Econome trica*, Oct. 1974, *42*, 1,093–1,113.

Morris Perlman, "The Roles of Money in a Economy and the Optimum Quantity c Money," *Economica*, Aug. 1971, *38* 233–52.

Roy Radner, "Competitive Equilibrium unde Uncertainty," *Econometrica*, Jan. 1968, *36* 31–58.

George J. Stigler, "The Economics of Infor mation," *J. Polit. Econ.*, Mar. 1961, *69* 213–25.

A. M. Ulph and D. T. Ulph, "Transaction Costs in General Equilibrium Theory— Survey," *Economica*, Nov. 1975, *42* 355–72.

E. C. H. Veendorp, "General Equilibriur Theory for a Barter Economy," *Wester Econ. J.* (now *Economic Inquiry*), Ma 1970, *8*, 1–23.

Price Expectations and Stability in a Short-Run Multi-Asset Macro Model

By EDWIN BURMEISTER AND STEPHEN J. TURNOVSKY*

Traditionally, short-run macroeconomic models are void of asset price dynamics. One explanation of this neglect involves the notion of "short-run equilibrium." If one *defines* a short-run equilibrium to be a situation in which relative asset prices are constant, the problem of short-run asset price dynamics is assumed away at the outset.

But this abstraction from the dynamics of relative asset prices in general is unjustified. The typical short-run macroeconomic equilibrium is characterized by two features:

1) The stocks of assets (money, bonds, and capital goods) are fixed (instantaneously).

2) Asset markets are in equilibrium in the sense that we observe points lying on the asset demand functions.

In general neither 1) nor 2) requires that relative asset prices be constant in short-run equilibrium But as we shall note below, one case in which this *will* be so is if the actual prevailing values of the real rates of return and coupon rates on the assets are expected to continue unchanged throughout all future periods. Under this extremely restrictive assumption of static expectations the existing asset prices will be expected to continue throughout the future, and the price dynamics degenerates. More generally we define *short-run dynamic equilibrium* to be a situation in which 1) and 2) are satisfied and in addition relative asset prices are constant.

The uniqueness and stability of such a short-run dynamic equilibrium becomes the crucial question that we will discuss below. If we discover that an economy always converges rapidly to a unique short-run dynamic equilibrium, that fact would provide a justification for models which do not contain short-run price dynamics. On the other hand, rapid convergence cannot be presumed a priori, and in fact we shall discover that some rather strong regularity conditions are required to ensure stability. When these regularity conditions are not satisfied, the possibility arises of short-run dynamic *instability*, a result that would destroy the validity of macroeconomic models formulated on the assumption that a short-run dynamic equilibrium prevails at every instant. Thus we will carefully examine the underlying mechanism generating price dynamics. Since we do not postulate dynamic equilibrium, our formulation may be viewed as a "disequilibrium approach" to short-run macroeconomics.

Our model contains three main ingredients:

(i) A new adaptive-type mechanism for generating price expectations simultaneously with a portfolio rate of return condition.

(ii) A savings function, postulated to depend upon (expected) disposable income and wealth, with disposable income defined to ensure the consistency of planned savings and planned (expected) wealth accumulation.

(iii) Asset demand functions that determine the value of the stocks of various assets as functions of expected rates of return, income, and wealth. Capital gains play a crucial role because they enter via the definition of the expected rates of return on assets.

*The authors are Professors of Economics at the University of Virginia and the Australian National University, respectively Burmeister wishes to acknowledge with thanks research support provided by the National Science Foundation and the Center for Advanced Studies at the University of Virginia

We will introduce (i)-(iii) in Sections I–III below; we then combine these features and briefly describe the behavior of the complete model.

I. Price Expectations and the Rates of Return

We consider a model with $n + 1$ assets, letting the $(n + 1) - st$ asset be the *numéraire* good, money. The stocks and prices of these assets at time t are $(X_1(t), \ldots, X_n(t), X_{n+1}(t))$ and $(P_1(t), \ldots, P_n(t), P_{n+1}(t))$, respectively, where $P_{n+1}(t) = 1$, so *wealth*, in terms of the *numéraire* (money) is

$$(1) \qquad W(t) = \sum_{i=1}^{n+1} P_i(t)X_i(t)$$

The stocks $X_i(t)$ are assumed to remain fixed at some arbitrary levels $\bar{X}_i$. Prices, however, will in general vary over time, with a major part of the dynamics being due to the evolution of price expectations.

We suppose that the price expectations for asset i satisfy the mixed total-partial differential equation

$$(2) \quad \dot{\pi}_i(t) = \beta_i \pi_{i,2}(t,t), \qquad \beta_i = constant > 0,$$

where $\pi_i(s,t)$ denotes the expectation formed at time t for the asset price $P_i(s)$, at time $s \geqq t$, and where

$$\pi_{i,1}(t,t) \equiv \left.\frac{\partial \pi_i(s,t)}{\partial s}\right|_{s=t}, \qquad \pi_{i,2}(t,t) \equiv \left.\frac{\partial \pi_i(s,t)}{\partial t}\right|_{s=t},$$

and

$$\dot{\pi}_i(t) \equiv \frac{d}{dt}\pi_i(t,t) = \pi_{i,1}(t,t) + \pi_{i,2}(t,t).$$

This equation has been derived elsewhere as the limit of a discrete time adaptive-type expected price mechanism, and space precludes a complete discussion here; see Burmeister and Turnovsky (1976a). A crucial feature of our expected price mechanism is that it satisfies what we have called the *weak consistency axiom*

$$(3) \qquad \pi_i(t,t) = P_i(t).$$

That is, the expectation formed at time t, for the price expected to prevail at that same instant of time t, must equal the actual price prevailing at that time. Underlying this axiom is the assumption that forecasters have instantaneous and costless access to information, so that they learn $P_i(t)$ the moment it occurs. Under these conditions any rational forecasting process must satisfy (3); if such information is not available instantaneously or is costly, then (3) need not hold. It is also important to realize that expectations generated by our mechanism may be completely rational in the sense that given the information available to economic agents, they are the maximum likelihood estimates of price levels; see for example, Benjamin Friedman. As is conventionally assumed, we take $\pi_i(s,t)$ to be point estimates of means held with subjective certainty. Ideally an uncertainty model would involve the entire probability distribution of the predicted variables, but this would be intractable for our purposes. As formulated, our model is in principle capable of empirical verification.

The net real rate of return on asset i is obtained as follows. The current price of a unit of that asset (in terms of the *numéraire*) must equal the discounted present value of the expected future earnings of that asset. Likewise, the expected price of asset i for time $(t + h)$, say, formed at time t, must equal the expected discounted earnings of that asset from time $t + h$ on. Assuming that expected dividends, (upon which the expected future earnings are based), and expected rates of return, satisfy weak consistency axioms analogous to (3) above, and taking the partial derivative with respect to h, we can show that the instantaneous rate of return on asset i at time t is given by

$$(4) \qquad r_i(t) = \frac{w_i(t)}{P_i(t)} + \frac{\pi_{i,1}(t,t)}{P_i(t)}.$$

The term $w_i(t)$ is the "dividend" or "coupon flow" that the ith asset pays at time t (expressed in terms of money). Thus the instantaneous rate of return $r_i(t)$ consists of the coupon rate of return $w_i(t)/P_i(t)$, plus the expected rate of capital gain $\pi_{i,1}(t,t)/P_i(t)$; see Burmeister and Turnovsky (1976a), Burmeister and Daniel Graham, and Turnovsky. In the case of the *numéraire*, money, $\pi_{n+1,1}(t,t) = 0$ so that adding the postulate that money yields no return ($w_{n+1} = 0$), we have $r_{n+1}(t) = 0$.

Note that if one makes the simplifying assumption that the current values of $w_i(t)$ and $r_i(t)$ are expected to prevail indefinitely, equation (4) simplifies to

$$(4a) \qquad r_i(t) = \frac{w_i(t)}{P_i(t)}.$$

Equation (4a) is in fact the relationship between asset price and return typically adopted in current macroeconomic theory; see, for example, James Tobin It holds under the assumption of static expectations, in which case the current asset price is also expected to prevail indefinitely; expected capital gains are zero and the price dynamics collapses

Combining equations (2), (3), and (4) yields

$$\pi_i(t) = \pi_{i,1}(t,t) + \pi_{i,2}(t,t) = \frac{\beta_i}{1-\beta_i}[w_i(t) - r_i(t)P_i(t)]$$

or, using the consistency axiom (3)

$$(5) \qquad \dot{P}_i = \pi_i = \psi_i[w_i - r_iP_i], \quad i = 1, \ldots, n,$$

where $\psi_i \equiv \frac{\beta_i}{1-\beta_i}$ and we assume $0 < \beta_i < 1$. The time variable t has been dropped to simplify notation. Henceforth we also shall assume that all the coupon rates w_i are exogenously given and constant.

Equations (5) play a fundamental role and have been derived in complete detail by Burmeister and Turnovsky (1976a). Loosely speaking, the expectations equation (2) is "backward looking" in time, while the rate of return equation (4) is "forward looking" in time. Combining these two effects yields a total time derivative for expected and actual asset prices in continuous time. *Thus equations (5) determine the rate of change of prices as the simultaneous solution to expectations equations and rate of return equations*

II. Savings and the Flow Constraint

Recent work on the formulation of continuous time macroeconomic models has shown how such models must satisfy a flow, as well as a stock, constraint at each point of time; see Josef May and Stephen Turnovsky. In this model stock the constraint is:

$$(6) \qquad W^d(t,t) = \sum_{i=1}^{n+1} P_i\bar{X}_i = W(t).$$

The flow constraint that the planned rate of wealth accumulation must equal the planned rate of savings is

$$(7) \qquad s(t) = W_1^d(t,t) = \sum_{i=1}^{n+1} \pi_{i,1}(t,t)\bar{X}_i,$$

and (7) holds *if, and only if*, we define disposable income by

$$(8) \qquad y^D(t) \equiv y(t) + \sum_{i=1}^{n+1} P_i r_i \bar{X}_i$$

To proceed further we must specify the savings function which we take to depend upon disposable income and wealth:

$$(9) \qquad s = s(y^D, W), \quad 1 > s_1 > 0, s_2 < 0,$$

with y^D defined by (8). This relationship is in nominal terms. Thus the level of income is one of the endogenous variables in our model. Because the stocks of all assets, including productive capital goods, are fixed in our short run, y is proportional to money wage income. Hence if the money wage rate is also fixed, y and employment are proportional; our model determines short-run employment in the usual Keynesian fashion

Finally, substituting for $\pi_{i,1}(t,t)$ from the rate of return equations (4), and using (6), (8), and (9), the flow constraint (7) can be written as

$$(7a)\quad s\left(y+\sum_{i=1}^{n+1}P_i r_i \bar{X}_i,\ \sum_{i=1}^{n+1}P_i \bar{X}_i\right)=\sum_{i=1}^{n+1}(P_i r_i - w_i)\dot{X}_i$$

and the latter in turn can be solved for y:

$$(10)\quad y=\phi(P_1,\ldots,P_n;\ r_1,\ldots,r_n)$$

where

$$(11a)\quad \partial y/\partial P_i = [r_i X_i(1-s_1)-s_2 X_1]/s_1 > 0$$

$$(11b)\quad \partial y/\partial r_i = (1-s_1)P_i X_i/s_1 > 0$$

In particular, we deduce that any increase in r_i must raise the level of income in order to generate sufficient savings to match the desired increase in the rate of wealth accumulation.

III. Asset Demand Functions and Assumptions

The asset demand functions are

$$(12)\quad P_i X_i = F^i(r_1,\ldots,r_n,y,W),\qquad i=1,\ldots,n,n+1,$$

where it will be recalled that y denotes the level of income measured in terms of the *numéraire* good, money. As well as requiring the usual "adding up" conditions to hold, the analysis, which is described in Section IV, is based on the following assumptions.

Assumption (A.1) (gross substitutes)

$$F_i^i > 0 \quad\text{and}\quad F_j^i \leqq 0 \quad\text{for}\quad i\neq j$$
$$i=1,\ldots,n,n+1,$$
$$j=1,\ldots,n.$$

(A.1) says that all assets are gross substitutes

Assumption (A.2) (demand for money)

$$F_y^{n+1}\geqq 0 \quad\text{and}\quad F_y^i \leqq 0,\qquad i=1,\ldots,n.$$

(A.2) asserts that the transaction demand for money F_y^{n+1} is financed (instantaneously) by reducing the holdings of all other assets.

Assumption (A.3) (wealth effects)

$$F_W^{n+1}>0 \quad\text{and}\quad F_W^i \geqq 0,\qquad i=1,\ldots,n.$$

This assures that wealth has a positive effect on the demand for money and a non-negative effect on all other assets.

Assumption (A.4) (weak transactions demand for money)

$$F_i^{n+1}+\gamma P_i X_i F_y^{n+1}<0,\qquad i=1,\ldots,n,$$

where

$$\gamma=(1-s_1)/s_1>0.$$

Assumption (A.4) is rather strong, but it is crucial for our conclusions. It states that an increase in any yield r_i decreases the demand for money sufficiently to outweigh any increase in the transactions demand for money arising from the associated increase in income y which occurs in order to maintain the flow constraint, see (11b). Assumption (A.4) does appear empirically realistic

IV. Behavior of the Model

Space limitations prohibit a complete analysis of the model, but Burmeister and Turnovsky (1976b) contains proofs of the theorems stated below and is available from the authors upon request. The first result is that, given assumptions (A.1)–(A.4), equations (7a), (12), and (6) can be solved globally for

$$(13)\quad r_i=f^i(P_1,\ldots,P_n),\qquad i=1,\ldots,n.$$

The economic interpretation of this result is that equations (12) and (6) define a generalized "*LM*" curve, while equation (7a) defines a gen-

eralized "IS" curve. These generalized "LM" and "IS" curves enable us to solve for the rates of return, defined to include expected capital gains, as functions of the prevailing asset prices.

Equations (13) are now substituted into the dynamic price equations (5) to yield an autonomous system of differential equations in asset prices:

$$(14) \quad \dot{P}_i = \psi_i[w_i - P_i f^i(P_1, \ldots, P_n)], \quad i = 1, \ldots, n.$$

A *short-run dynamic equilibrium* is a rest point $(P_1^*, \ldots, P_n^*) > 0$ to (14) where $\dot{P}_i = 0$. We shall suppose that the exogenous dividends—the w_i's—are such that a rest point exists. The stationary values therefore satisfy the static asset pricing relationship (4a) Moreover, given our assumptions any rest point must be *unique*, and the remaining question is the stability of equilibrium

The strongest stability results obtain when there are three assets: asset 1 or "capital," asset 2 or "bonds," and asset $3 = n + 1 = 2 + 1$ or "money." This three asset $n = 2$ case corresponds to traditional macroeconomic models, and without additional assumptions we have proved that any rest point solution to the dynamic price equations (14) is locally asymptotically stable, independent of the adjustment speeds ψ_i.

When the number of assets exceeds three ($n > 2$), the same stability result can be proved provided we also assume.

Assumption (A.5)

The matrix $\left[\dfrac{\partial r_i}{\partial P_j}\right]$ has a positive diagonal and negative off-diagonal elements.

It can be proved that this sign pattern does prevail when income does not affect the demand for assets and when all cross effects in the asset demand functions are zero. Therefore assumption (A.5) may be satisfied when these conditions hold approximately, and its validity becomes an empirical question.[1]

V. Conclusions and Future Research

The formulation of consistent dynamic macroeconomic models is an important issue and this paper makes only modest progress. While we have obtained conditions under which a short-run dynamic equilibrium will be unique and stable, irrespective of the speed of adjustment, other issues requiring further investigation are also raised. The conditions under which stability have been obtained are stringent; to what extent can they be relaxed? Even if the system is stable, is the convergence of asset prices to equilibrium sufficiently rapid to justify the usual neglect of price dynamics? The answer to this question will depend upon the magnitudes of the adjustment coefficients, and will also involve analyzing the asset demand and savings functions. We have also assumed that the stocks of assets remain constant. It is most important to allow these to vary and to embed the dynamics of asset prices into a complete dynamic macroeconomic model which incorporates the accumulation of assets. The fact that a rigorous examination of short-run price dynamics is difficult is not excuse for its complete neglect without careful justification, and all macroeconomic models which ignore price dynamics may be seriously misleading if in fact a properly formulated model is featured by either instability or slow convergence to short-run dynamic equilibrium.

[1]The economic reason that assumption (A.5) implies stability may be briefly sketched as follows An increase in the price of asset one in terms of money, the *numéraire*, increases the volume of the stock of that asset more than it increases demand through wealth effects Thus for equilibrium to be restored in the market for asset one, its demand must be increased further through forcing up its real rate of return Likewise the additional wealth generated by a higher P_1 increases the demand for all other assets; given their fixed supplies, all their real rates must fall to restore equilibrium in these other asset markets

REFERENCES

Edwin Burmeister and D. Graham, "Multi-Sector Economic Models with Continuous Adaptive Expectations," *Rev. Econ. Stud.*, 1974, *16*, 323–36.

______ **and Stephen J. Turnovsky,** "The Specification of Adaptive Expectations in Continuous Time Dynamic Economic Models," *Econometrica*, 1976a, *44*, 879–905.

______ **and** ______, "Proofs for *AEA* Paper," Discussion Paper, 1976b.

Benjamin M. Friedman, "Rational Expectations are Really Adaptive After All," Discussion Paper No 430, Harvard Institute of Economic Research, Harvard University, Aug. 1975.

J. May, "Period Analysis and Continuous Analysis in Patkin's Macroeconomic Model," *J. Econ. Theory*, 1970, *2*, 1–9.

James Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit and Banking*, 1969, *1*, 15–29.

Stephen J. Turnovsky, "On the Formulation of Continuous Time Macroeconomic Models with Asset Accumulation," *Intern. Econ. Rev.*, *18*, 1977, forthcoming.

WELFARE ECONOMICS

Extended Sympathy and the Possibility of Social Choice

By KENNETH J. ARROW*

The new results in social choice theory that I wish to sketch here today were developed independently by two young scholars, Steven Strasnick, in an unpublished doctoral dissertation in philosophy at Harvard, and Peter J. Hammond, an English economist. The works have already been influential in manuscript and have in particular led to an excellent synthesis by the Belgian economists, Claude d'Aspremont and Louis Gevers, and it is their exposition which I shall largely follow. Most of this paper will be devoted to an exposition of the formal theory, though I shall omit proofs; at the end, I will make some comments on interpretation.

I. The Invariance of Social Choice Under Transformations of Utilities

In my book, I used the preference orderings of individuals over social states as the variables which determined the social orderings. As is well known, an alternative equivalent representation of individual preference would have been a real-valued utility function over social states, which, however, would have meaning only up to a monotone transformation.[1] Since no interpersonal comparisons appeared in my approach, the utility functions of the different individuals could be subject to independent monotone transformations.

Let,

N = set of individuals,

X = set of possible social alternatives.

Any given feasible set of alternatives is, then, a subset of X.

A utility function, $u(x)$, is a real-valued function on X. It defines an ordering over X in the usual way, x is preferred to y if and only if $u(x) > u(y)$. A specification of utility functions, one for each individual, $u_i(x)$ ($i = 1, \ldots, n$), can also be regarded as a single function $u(x, i)$ over the Cartesian product $X \times N$. Let,

U = set of real-valued functions on $X \times N$.

From now on, the term, "utility function," will refer to any member of U In these terms, a social welfare function or constitution (to use my now-favorite term) can be defined:

Definition: A *constitution* is a function, f, mapping U into orderings of X.

That is, we associate to each utility function (in the present sense, i.e., a utility function for each individual) a social ordering of social states.

Let u be a utility function, and let g_i ($i = 1, \ldots, n$) be n (strictly) monotone functions from real numbers to real numbers. Let a utility function, u', be defined by,

$$u'(x, i) = g_i\,[u(x, i)].$$

If we maintain the strictly ordinal approach of Arrow and also prohibit interpersonal comparisons, then there would be no operational distinction between u and u'; the only evidence we have is the ordering of social states for each individual, and that is the same for u and u'. Hence, this ascetic viewpoint would require

*Harvard University

[1]Strictly speaking, not every preference ordering can be represented by a real-valued indicator, but this restriction can be neglected here It is no restriction if the number of alternatives is finite

that the social orderings defined by u and u' be the same.

Ordinal Invariance: If there exist strictly monotonic functions g_i $(i = 1, \ldots, n)$ from real numbers to real numbers such that,

$u'(x, i) = g_i[u(x,i)]$ for all x and i,
then $f(u) = f(u')$.

As is well known (see Arrow, 1963), Ordinal Invariance, together with the other assumptions usually made about the constitution (see next section), implies that there exists no constitution.

The question then arises, are there less demanding forms of invariance which permit the existence of satisfactory constitutions and which can be justified in terms of actual or at least hypothetical observations.

We will admit the meaningfulness of interpersonal ordinal comparisons. That is, we regard as meaningful statements of the form,

individual i in state x is better off than individual j in state y.

Whatever one may think of interpersonal comparisons, at least these are ordinal and therefore may be interpreted as hypothetical choice. I defer more detailed defense and critique to Section V.

In this case, $u(x, i)$ may be interpreted as the utility derived by individual i in state x Statements of the form,

$$u(x, i) > u(y, j),$$

are to be preserved under transformations. Thus, the permitted transformations are monotone transformations of the whole utility function, but not transformations which differ from individual to individual.

Co-ordinal Invariance: If there exists a strictly monotone function, g from real numbers to real numbers such that,

$u'(x, i) = g[(x, i)]$, all x and i,
then $f(u) = f(u')$.

II. Other Conditions on Social Choice

We retain conditions on social choice like those in Arrow (1963), though restated to be compatible with the present definition of a constitution as a mapping from utility functions rather than from n-tuples or orderings.

Since, for each u, $f(u)$ is an ordering, the notation,

$$x\, f(u)\, y,$$

means,

x is at least as good as y in the social ordering induced by the utility function u.

The strict preference ordering defined by u will be denoted by $f^P(u)$.

The (controversial) assumption of independence of irrelevant alternatives will be stated here only for preferences, i.e , for choice from two-member sets. Only that part of the assumption was used in Arrow (1963).

Binary Relevancy: If u and u' are such that, $u(x, i) = u'(x, i)$ and $u(y, i) = u'(y, i)$, all i, then $xf(u)\, y$ if and only if $xf(u')\, y$.

The democratic condition that all individuals are to count equally will be here represented in the strong form of symmetry among individuals, instead of the very weak nondictatorship condition.

Anonymity: Let σ be a permutation of N. If, $u(x, i) = u'[x,\sigma(i)]$ for all x and i, then $f(u) = f(u')$.

The Pareto condition used in Arrow (1963) is the weak condition,

Weak Pareto: If $u(x, i) > u(y, i)$, all i, then x $f^P(u)\, y$.

For certain purposes, we will want the stronger condition that if at least one individual is better off by a change, while none are hurt, the change should be made.

Strong Pareto: If $u(x, i) \geqq u(y, i)$, all i, and, for some j, $u(x, j) > u(y, j)$, then $x f^P (u) y$.

Finally, there is an interesting generalization of the Pareto condition, an implication of which has interesting consequences. Up to this point, we have taken the range of individuals, N, as given. But suppose we assume that we have a constitution for every set of individuals. We would expect to have some consistency conditions among these constitutions. Indeed, the Pareto principle is one such. If there is only one individual in the world, the social ordering is simply his. Then the Weak Pareto principle can be interpreted as asserting that if all one-individual subsets prefer x to y, then so does the whole society; while the Strong Pareto principle says that if all one-individual subsets weakly prefer x to y, while at least one has a strong preference, then society prefers x to y. The one-individual subsets are a partition of the entire set of voters. Then it is reasonable to extend to principle to cover all partitions of the voters; for each subset, we are assuming that the constitution prescribes a mapping of the utility function (restricted to the individuals in that subset) into a social ordering.

Since N is now variable, we define,

U_N = set of real-valued functions on $X \times N$.

A constitution for any given sets of voters N is now supposed to define social orderings for utility functions on every subset of voters.

Definition: A constitution is a family of functions, f_N, defined for all sets of voters N, mapping U_N into orderings of X.

If $u \in U_N$, some N, and $M \subset N$, then u_M will be the function u restricted to individuals in M and so belongs to U_M. Hence, for any $u \in U_N$, $f_M(u_M)$ is defined for every $M \subset N$.

Generalized Strong Pareto: (a) If N consists of the single individual, i, $f_N(u)$ is the ordering defined by the utility indicator $u(x, i)$. (b) If Q is a partition of N, and, for all $M \in Q$, $x f_M(u_M) y$, while for some $M' \in Q$, $x f^P_{M'}(u_{M'}) y$, then $x f^P_N(u) y$.

This principle was stated by Strasnick. It had been used earlier by Young (p. 44).

Here we use the generalized Strong Pareto principle only in the special form of,

Elimination of Indifferent Individuals: Let N be partitioned into M' and M''. Suppose $u \in U_N$, $u' \in U_N$, and $u_{M'} = u'_{M'}$, while, for all x and y, $u(x, i) = u(y, i)$ and $u'(x, i) = u'(y, i)$ for all $i \in M''$. Then $f_N(u) = f_N(u')$.

III. The Theorems

We know of course that Ordinal Invariance, Binary Relevancy, Anonymity, and the Weak Pareto condition are incompatible. If, however, we replace Ordinal Invariance by Co-ordinal Invariance, the conditions are indeed satisfied and, in fact, by the maximin principle, i.e.,

$$x f(u) y \text{ if and only if } \min_i u(x, i) \geqq \min_i u(y, i).$$

This condition also satisfies Elimination of Indifferent Individuals and the Generalized Weak Pareto principle (an obvious analogue of the Generalized Strong Pareto principle given in the preceding section). It does not satisfy the Strong Pareto principle; however, as Sen (p. 138, fn. 11) has noted, a simple modification will lead to satisfaction.

Lexical Maximin Principle: For any alternative x and utility function u, rank the individuals in increasing order of $u(x, i)$, and let $i(x, k)$ be the kth ranking individual; ties can be broken arbitrarily. For any pair of alternatives x, y, let

$$k(x, y) = \min \{k \mid u[x, i(x, k)] \neq u[y, i(y, k)]\},$$

if defined. Then $x f^P (u) y$ if and only if $u[x, i(x, k(x, y))] > u[y, i(y, k(x, y))]$

The Lexical Maximin Principle satisfies Co-ordinal Invariance, Binary Relevancy, Anonymity, and the Generalized Strong Pareto condition.

It is not, however, the only principle satisfying these conditions. Indeed, the maximax principle,

$$x f(u)\, y \text{ if and only if } \max_i u(x, i) \geqq \max_i u(y, i),$$

also satisfies Co-ordinal Invariance, Binary, Relevancy, Anonymity, and the Weak Pareto condition; and clearly, the Lexical Maximax principle, defined in the obvious way, satisfies all the conditions stated above for Lexical Minimax.

What is surprising is that these are the only two such conditions; and, by making a very weak equity assumption, the Lexical Maximax principle can be ruled out.

We define two more conditions; they will not be regarded as primary, but their relations with the other conditions will be stated.

Strong Equity: For all $u \varepsilon U_N$, all x and y in X, and all i and j in N, if $u(x, g) = u(y, g)$ for $g \neq i, j$, and $u(y, i) < u(x, i) < u(x, j) < u(y, j)$, then $x f(u)\, y$.

I.e., if all but two individuals are indifferent between x and y, and one individual is better off than the other in both x and y, his choice should not be binding. By itself, this amounts to putting a weak version of Rawls right into the axiom system.

The dual assumption to Strong Equity is,

Inequity: For all $u \varepsilon U_N$, all x and y in X, and all i and j in N, if $u(x, g) = u(y, g)$ for $g \neq i, j$, and $u(y, i) < u(x, i) < u(x, j) < u(y, j)$, then $y f^P (u)\, x$.

The better-off individual *always* prevails.

THEOREM 1: If the constitution satisfies Binary Relevancy, Anonymity, Co-ordinal Invariance, and Elimination of Indifferent Individuals, then either Strong Equity or Inequity holds.

This result may seem surprisingly strong, and its proof takes numerous steps. However, an intuitive sketch can be given for the case of two individuals. In the first place, the assumptions of Binary Relevancy, Anonymity, and Elimination of Indifferent Individuals can easily be shown to imply,

Neutrality: If σ is a permutation of X, and $u(x, i) = u'[\sigma(x), i]$ for all x and i, then $x f(u)\, y$ if and only if $\sigma(x)\, f(u')\, \sigma(y)$.

Changing the names of the alternatives does not matter. Suppose, then, both Strong Equity and Inequity fail. The failure of Strong Equity implies the existence of u, x, y, i and j, such that,

$$u(y, i) < u(x, i) < u(x, j) < u(y, j), \quad y f^P (u)\, x.$$

The failure of Inequity implies the existence of u', x', y', i', and j' such that,

$$u'(y', i') < u'(x', i') < u'(x', j') < u'(y', j'), \quad x' f(u')\, y'.$$

But from Neutrality and Anonymity, we can take $x' = x$, $y' = y$, $i' = i$, $j' = j$. Then on the set consisting of the four elements, (x, i), (x, j), (y, i), and (y, j), u and u' give the same ordering. By Binary Relevancy, the ordering over other elements is irrelevant to the choice between x and y. But an ordinal transformation common to the two individuals takes u into u', in contradiction to Co-ordinal Invariance.

The extension to many individuals is laborious but relies mainly on Elimination of Indifferent Individuals.

On the other hand, as Hammond and Strasnick have shown, Strong Equity with the other assumptions implies Lexical Maximin.

THEOREM 2: If the constitution satisfies Binary Relevancy, Anonymity, Co-ordinal Invariance, the Strong Pareto principle, and Strong Equity, then it is the Lexical Maximin principle.

The Strong Equity assumption postulates the result for two individuals; the problem in the proof is to extend it to any number of individuals.

Of course, entirely dual to Theorem 2, we have,

THEOREM 3: If the constitution satisfies Binary Relevancy, Anonymity, Co-ordinal Invariance, the Strong Pareto principle, and Inequity, then it is the Lexical Maximax principle.

If we add the Strong Pareto principle to the assumptions of Theorem 1, then Theorems 1, 2 and 3 together tell us we have reduced the range of possible constitutions to two, the Lexical Maximin and the Lexical Maximax. To eliminate the second, it is sufficient to deny the Inequity assumption, instead of imposing the apparently stronger Strong Equity condition. An assumption which contradicts Inequity is,

Minimal Equity: There exist $u \in U_N$, $x \in X$, $y \in X$, and $j \in N$, such that, for all $i \neq j$, $u(y, i) < u(x, i) < u(x, j) < u(y, j)$, and $x\, f(u)\, y$.

That is, there is at least one utility function and one individual, such that the individual is better off than any one else under either alternative and has preferences opposite to all of theirs, and the given individual does not prevail. Since Lexical Maximax clearly does not satisfy Minimal Equity, it is clearly from Theorem 3 that Minimal Equity contradicts Inequity. From Theorems 1 and 2, then, we must have Lexical Maximin.

THEOREM 4: If the constitution satisfies Binary Relevancy, Anonymity, Co-ordinal Invariance, the Strong Pareto principle, and Minimal Equity, then it is the Lexical Maximin principle.

IV. Evaluation of the Axiomatic Justification of Maximin

Suppose we assume for a moment the meaningfulness of interpersonal ordinal comparisons (see next section). Do we find the results convincing?

There are two reservations which come to mind. The first is narrower, that is, it accepts virtually all of the argument. The whole argument is basically symmetrical in "best-off" and "worst-off" individuals.[2] These two are indeed distinguished from the others; that is basically an implication of Co-ordinal Invariance.[3] But to exclude letting decisions be made in the interests of the best-off requires some form of direct assumption to the contrary, if only in the weak form of the Minimal Equity assumption.

A second reservation can be put as a continuity requirement. Suppose a change from x to y diminishes the utility of the worst-off by some very small amount but increases the utility of all others by a great deal. Surely it seems reasonable to argue that if the loss to the worst-off is small enough and the gain to everyone else large enough, society should prefer y to x. Indeed, if there were *no* loss to the worst-off and a gain to all others, the Lexical Maximin rule would call y a strict improvement; hence, by any kind of continuity argument, the preference for y over x should be maintained if the utility loss to the worst-off is sufficiently small.

Thus, adding a continuity requirement to the hypotheses of Theorem 4 leads to an impossibility theorem.

This is by no means a "formal" matter. Clearly, the intuition behind the continuity requirement is a small step in the direction of utilitarian ethics; even the worst-off member of the society might be made to suffer if there is enough benefit to others.

There is one striking case, of great practical importance, where our intuition is in favor of

[2]Robert Nozick has stressed this point to me

[3]The special role of the extremes has appeared in a parallel context, that of decision making under uncertainty, where we do not wish to ascribe probabilities (directly or indirectly) to the possible states of the world The states of the world correspond to individuals, the actions to be chosen to alternative social states. Certain approaches imply co-ordinal invariance. See J Milnor, Theorem 4 and Arrow and Leonid Hurwicz

utilitarianism in some form as against any minimax rule. I refer to allocation over time. Typically, we expect future generations to be better off than we are. Should we save for them either directly or in the form of public investments? A maximin rule would surely say no. But if investment is productive, so that, in terms of goods, the next generation gains more than we lose, we usually feel that some investment is worthwhile even though the recipients will be better off than we are.[4]

V. The Operational Significance of Interpersonal Ordinal Comparisons

The discussion of the last section suggests that, if anything, even interpersonal ordinal comparisons are not sufficient to take account of our intuitions of justice as derived from a social choice framework The requirement of Co-ordinal Invariance may still be too strict. In this section, I will address the opposite question; whether such comparisons have any meaning, i.e., whether the invariance criterion should not perhaps be even stricter.

The possibility of such comparisons has already been defended in different forms by P. Suppes, S. C. Kolm, Part C, and myself (1963, pp. 114–15). I will comment only briefly, to avoid repeating my earlier arguments excessively.

The concept of a preference ordering or, by extension, of a utility function is related to hypothetical choices. Its usual use is in a complete theory, say of individual behavior, in which the preference ordering and the feasible set jointly determine the chosen alternative. The preference ordering is thought of as given before the feasible set is known and therefore determines choices among all possible pairs of alternatives. The feasible set prescribes which alternatives are in fact available. It makes sense, therefore, to include in our information choices which are not in fact feasible though they are conceivable.

Now we can say that among the characteristics which determine an individual's satisfaction are some which are not, at least at the moment, alterable. An individual who is ill can meaningfully be said to prefer being well. If in fact there were some medical means of cure, we would test this preference by asking if he will purchase the services. But clearly the preference would be there whether or not medicine was useful.

We may suppose that everything which determines an individual's satisfaction is included in the list of goods. Thus, not only the wine but the ability to enjoy and discriminate are included among goods. It is, in fact, true that only some of the goods so defined are transferable among individuals while others are not But that consideration enters into the definition of the feasible set, not that of the ordering. If we use this complete list, then everyone should have the same utility function for what he gets out of the social state. This does not, of course, mean that individuals agree on the utility of a social state, since what they receive from a given state differs among individuals.

Formally, we may suppose a space, Y, which defines the range of possible implications of a social state for an individual. Since the state defines, for each individual, everything that characterizes his satisfactions, the space Y is the same for all individuals. It includes goods, tastes, and the reactions of others to the extent that individuals care about each other. All individuals have the same preferences over Y. Let $u(y)$ be the ordinally defined utility indicator. Each state in X defines implications for every individual. Let $G_i(x)$ $(i = 1, \ldots, n)$ be for each individual a mapping from X to Y, expressing these implications Then we can identify,

$$u(x, i) = u[G_i(x)],$$

and this possesses Co-ordinal Invariance.

This is at least one way to interpret and defend interpersonal ordinal comparisons. (There are others.)

[4]I do not find Rawls's theory of just saving (1971, section 44) at all clear; it seems to avoid a rigid maximin rule without providing a clear substitute. See Arrow (1973) and P Dasgupta

I cannot, however, conclude without admitting some difficulties. I can think of two, though perhaps they are the same looked at somewhat differently. For one thing, if your satisfaction depends on some inner qualities that I do not possess, then I really have not had the experience which will enable me to judge the satisfaction one would derive from that quality in association with some distribution of goods. Hence, my judgment has a probability element in it and therefore will not agree with your judgment. But it is essential to the present construction that the comparisons of individual i in state x with individual j in state y be the same whether the comparison is made by i, j, or a third individual, k.

The second difficulty is that reducing an individual to a specified list of qualities is denying his individuality in a deep sense. In a way that I cannot articulate well and am none too sure about defending, the autonomy of individuals, an element of mutual incommensurability among people seems denied by the possibility of interpersonal comparisons. No doubt it is some such feeling as this that has made me so reluctant to shift from pure ordinalism, despite my desire to seek a basis for a theory of justice.

REFERENCES

Kenneth J. Arrow, *Social Choice and Individual Values*, 2nd Ed., New York and New Haven 1963.

———, "Rawls's Principle of Just Saving," *Swedish J. Econ.*, 1973, *75*, 323–335.

——— **and L. Hurwicz,** "An Optimality Criterion for Decision-Making Under Ignorance," in C. F. Carter and J. L. Ford (eds.), *Uncertainty and Expectations in Economics*, Oxford 1972.

P. Dasgupta, "On Some Alternative Criteria for Justice Between Generations," *J. Pub. Econ.*, 1974, *3*, 405–23.

C. d'Aspremont and L. Gevers, "Equity and the International Basis of Collective Choice," *Rev. Econ. Stud.*, forthcoming.

Peter J. Hammond, "Equity, Arrow's Conditions, and Rawls' Difference Principle," *Econometrica*, 1976, *44*, 793–804.

S. C. Kolm, *Justice et Équité*, Paris 1972.

J. Milnor, "Games against Nature," in R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), *Decision Processes*, New York 1954.

Amartya K. Sen, *Collective Choice and Social Welfare*, San Francisco 1970.

S. L. Strasnick, *Preference Priority and the Maximization of Social Welfare*, Doctoral dissertation, Harvard University 1975.

P. Suppes, "Some Formal Models of Grading Principles," *Synthese*, Dec. 1966, *16*, 284–306.

H. P. Young, "An Axiomization of Borda's Rule," *J. Econ. Theory* Sept. 1974, *9*, 53–52

Information and Performance in the (New)² Welfare Economics

By STANLEY REITER*

The title refers to a distinction between the new welfare economics, (now not so new) and a newer welfare economics. The objectives of the new welfare economics were to provide principles for evaluating and comparing alternative allocations in a given economy. The Pareto principle is as far as common agreement goes in this problem, and there are now dissenters even from that. Perhaps the main accomplishment of the new welfare economies was to derive conditions characterizing (Pareto) efficient production and exchange, and to show that in classical economies (those in which preferences and production possibilities have suitable convexity and continuity properties) the equilibria of the competitive mechanism precisely meet the conditions characterizing Pareto optimal production and exchange. That is, in classical economies the competitive equilibrium allocations and the Pareto optimal allocations are two names for the same collection of allocations These are, of course, the classical welfare theorems of Kenneth Arrow, Gerard Debreu and Tjalling Koopmans.

However, during the period of the development of this line of theory there was also interest in nonclassical economies and in systems other than the competitive one. People sought to characterize Pareto-optimal allocations when there are indivisibilities, increasing returns or externalities. The work of Harold Hotelling provides examples, and Pigou and Marshall also discussed such problems. Marginal cost pricing proposals, and the organizational proposals of the Lange-Lerner-Taylor type, were attempts to design economic systems which would function satisfactorily in nonclassical situations, where the competitive system lacks optimal properties. These proposals constituted a shift of the focus of welfare economics to the system of economic organization. As might be expected, this shift of focus brought new subjects into the discussion. Questions of administrative feasibility and the costs of operating the system were raised in connection with the early controversies over central planning. Questions were raised about the extent to which private incentives are in conflict with the system, thereby creating a need for policing with its attendant costs, or divergences of the outcome produced by the system from the one it was designed to produce. These lines of thought lead to the central problem of the (new)² welfare economics.[1]

That problem is not merely one of evaluating alternative allocations in a given economy, but of comparing the functioning of alternative systems operating in a class of economic environments, such as the classical ones, or alternatively those with indivisibilities or other nonconvexities. In such a problem the allocations which are the outcome of the system are just one of the important aspects of its functioning, but others are also important, among them its administrative feasibility, the costs of operating the system itself, the extent to which private incentives are incompatible with the system. These considerations pertain to properties of the economic mechanism itself, not merely of its resulting allocations. All these properties of an economic system must be weighed in order to evaluate the system and compare it with alternatives. For this a theory is

*Northwestern University I wish to thank Morton I Kamien, Peter McCabe and D. John Roberts for helpful comments. This research was partially supported by National Science Foundation grant SOC-7103784

[1] I have called it "(new)²" to suggest difference by an order of magnitude from the "new" welfare economies.

needed which encompasses these elements. This is the main objective of the (new)[2] welfare economics, to provide a normative theory of economic mechanisms.

It is with the publication of Leonid Hurwicz's paper "Informational Efficiency of Resource Allocation Mechanisms" in 1960 that a formal structure for the comparative study of economic mechanisms appears in economics. The fifteen or so years since then have seen a substantial development of this field. New results, new questions and new methods have accumulated. It is by now a thriving industry perhaps just emerged from its infancy and entering the phase of exponential growth. I shall try here to give some picture of its methods, questions and results so far.

An economic system, insofar as it determines the allocation of resources, may be viewed as a kind of machine which accepts as inputs the basic data of an economy and produces as its output an allocation of commodities among the participants in the economy.

The basic data of an economy, briefly an *economy* or *economic environment*, consists of the list of agents $\{1, \ldots, n\}$ the list of commodities $\{1, \ldots, \ell\}$ and the *characteristics* of each agent. These typically are, for agent i, his preference relation $\lesssim_i$, or its representation by a utility function u^i, his technology T^i, given, say, as a production possibilities set, and his initial endowment vector ω^i. Denote the characteristic of the ith agent by $e^i = (\lesssim_i, T^i, \omega^i)$ for $i = 1, \ldots, n$. We assume the commodity space to be the ℓ-dimensional Euclidean space R^ℓ, the same for all the economies we are considering, and that the list of participants is also the same for all economies. With these assumptions, an economy may be specified by the n-tuple of characteristics of the n agents; thus, $e = (e^1, \ldots, e^n)$ denotes an *economy* or economic *environment*. Note that this formulation does not exclude externalities.

Economic activity results in allocations which are representable by points in $R^{n\ell}$. An economic system generally must deal with more than one set of economic data, just as a computational algorithm is designed to accept more than just one numerical problem. So the class E of economies to be accepted by the mechanism must be specified.

The allocations (or trades) which are *feasible* for the economy e are denoted $\mathscr{F}(e)$. Thus, $\mathscr{F}$ is a *correspondence* (a multivalued function) which assigns to each economy e in the specified class E of economies the set of feasible allocations (trades) for e. For a two-person, two-good pure exchange economy e, the set of feasible allocations $\mathscr{F}(e)$ consists of the allocations given by the Edgeworth Box whose size corresponds to the total endowment of commodities.

Similarly, the set of allocations (or trades) of e considered *desirable* is also represented by a correspondence $\mathscr{P}$, which assigns to each economy e in E the subset $\mathscr{P}(e)$ of the feasible allocations $\mathscr{F}(e)$ which are considered desirable. Frequently $\mathscr{P}$ is taken to be the Pareto correspondence, in which case $\mathscr{P}(e)$ is the set of Pareto optimal allocations (trades) for e.[2]

An initial distribution of knowledge about the economy is assumed. Each agent knows something, but generally not everything, about the economy he is in. We assume here, as is typical in this literature, that each agent i knows directly only his own characteristic e^i. Since to determine whether or not an allocation is feasible in an economy e, let alone Pareto-optimal, in general requires data from all of the characteristics (e.g., the total initial endowment $\omega = \sum_{i=1}^{n} \omega^i$), no agent by himself knows enough to figure out the feasible allocations. Optimal coordination of economic activity in general requires communication among the agents when knowledge about the economy is dispersed. We discuss next how the process is modeled.

A resource allocation system is modeled in two closely related ways. The first, as an *adjustment process* (Hurwicz 1960) and the second as a *mechanism* (K. Mount and Reiter,

[2]A trade y is Pareto optimal in e if the allocation $x = \omega + y$, where ω is the initial allocation in e, is a Pareto optimal allocation in e.

1974a and 1974b). There are two stages. In the spirit of tatonnement agents first communicate; "real" economic action is taken only when "equilibrium" is reached. (Nontatonnement formulations can also be given.) Thus, in the first stage, the agents communicate with one another by sending formalized messages taken from a specified set of messages or *language*. After no further communication is worthwhile, or when this stage otherwise comes to an end, the final message is translated into action.

The adjustment process formulation models the iterative exchange of messages as follows. Agent i can emit a message $m^i(t)$ at time t which is drawn from his *language* M^i, $i = 1, \ldots, n$. He can select this message on the basis of what he knows at time t (just prior to the exchange of messages at t). His response at t is given by a function f^i according to the equation

$$(*)\ m^i(t) = f^i(m(t-1), e^i) \qquad i = 1, \ldots, n$$

where $m^i(t)$ is an element of M^i and $m(t) = (m^1(t), \ldots, m^n(t))$ is an element of the *message space* $M = M^1 \times \ldots \times M^n$. Thus, the process of communication is modeled by a system of temporally homogeneous first order equations in the messages, with the individual characteristics as parameters.[3] That the message of i depends only on e^i and not on any other economic data expresses the idea that agent i initially knows only e^i, and that the only way he has of acquiring additional information is via the communication process. This property of the *response function* f^i Hurwicz, (1972a) called *privacy*. (It was defined in Hurwicz 1960, but given another name.)

Using vector notation we may abbreviate the system (*), to

$$m(t) = f(m(t-1), e)$$

[3]This formulation covers the case of agents with finite memories, since a temporally homogeneous difference equations of finite order can be transformed into one of first order by making the message big enough. Instead of having two concepts of informational capacities: memory and message size, this formulation allows us to capture both in one

A joint message $\bar{m} = (\bar{m}^1, \ldots \bar{m}^n)$ in M is a *stationary message* for the economy e and the response function f if and only if

$$\bar{m} = f(\bar{m}, e).$$

We may assume that stationary messages exist for the economies and response functions we consider and that solutions of the difference equations converge to them.

Stationary messages are translated into *actions* or *outcomes* by the *outcome function h;* thus, if $\bar{m}$ is a stationary message for e, then $a = h(\bar{m})$ is an outcome or action determined by the mechanism for the economy e. If e is a pure exchange economy, then typically $h(\bar{m})$ would denote a vector of trades. Thus, an adjustment process is a triple (M, f, h) where M is the message space, f the (vector of) response function(s), and h the outcome function of the process. An adjustment process whose response functions satisfy privacy is said to *preserve privacy*

Such a mechanism can be represented in another somewhat more general way which is sometimes more convenient. Here we suppose communication takes place in one step, rather than iteratively, and that what is communicated is the collection of all joint messages "acceptable" to the agent. This can be thought of as a function which gives the message agent i would emit in response to the other components. To represent the process (M, f, h) in this way, we define the correspondence μ, called the (equilibrium) *message correspondence* by

$$\mu(e) = \{m \in M \mid f(m, e) - m = 0\}.$$

Thus, a message m belongs to the set $\mu(e)$ if and only if it is a stationary message of f at e. In order to give effect to the privacy requirement on f, the correspondence μ must have a special structure; it must be a *coordinate correspondence* (Mount and Reiter 1974a). Namely, there must exist correspondences μ^i for $i = 1, \ldots, n$ defined only for characteristics of i, and such that $\mu(e) = \bigcap^n \mu^i(e^i)$. Given the privacy

preserving process (μ, f, h) if we take μ^i to be given by

$$\mu^i(e^i) = \{m \in M \mid f^i(m, e^i) - m^i = 0\}$$

Then μ will also be privacy preserving. The privacy preserving process (M, f, h) can also be written in terms of the message correspondence as (M, μ, h).

The performance of such a mechanism can be represented in the following fundamental triangular diagram.

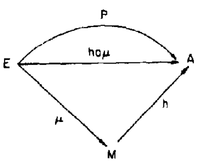


FIGURE 1

The message correspondence μ selects for each economy e in E a message (or messages) in M. These are translated by the outcome function h into outcomes in A. Application of the message correspondence μ to an economy followed by applying h to the messages $m = \mu(e)$, denoted $h \circ \mu$, associates the action(s) "computed by the mechanism to the given economy e in E, and thus can be considered to determine an arrow $h \circ \mu$ directly from E to A.

A process (M, μ, h) can be defined in two ways, one of which makes the composition $h \circ \mu$ a correspondence, the other a single-valued function. The definition of a process given by Hurwicz (1960) would make $h \circ \mu$ a correspondence, the definitions in Mount and Reiter (1974a) and (1974b) make it a function. In each case there is an appropriate definition of "$\mathcal{P}$-satisfactoriness" technically different but expressing the same concept. It amounts to this: a mechanism is $\mathcal{P}$-satisfactory on a class E of economies if and only if the outcomes produced by the mechanism exactly cover the correspondence $\mathcal{P}$, i.e., for each economy e in the class E every outcome is $\mathcal{P}$-optimal, and every $\mathcal{P}$-optimum is a possible outcome.

[1]Here and in what follows we take the mechanism to be defined so as to make $h \circ \mu$ a function.

A criterion of performance commonly used for mechanisms is that the relation between outcomes in A and economies should be the same as the one given by the Pareto correspondence. A mechanism is *Pareto satisfactory* on E if this condition is met. The classical welfare theorems (together with existence theorems) of Arrow, Debreu and Koopmans assert that the competitive mechanism is Pareto satisfactory on a class of economies satisfying certain convexity and continuity conditions. A weaker property sometimes studied is that for each economy e in E the outcome be Pareto-optimal,[2] i.e.,

$$h(\mu(e)) \in \mathcal{P}(e) \text{ for all } e \text{ in } E.$$

A mechanism with this property is called *nonwasteful* on E.

In terms of the diagram in Figure 1 we may ask whether there is a mechanism which for an arbitrary class of economies E, and an arbitrary performance criterian $\mathcal{P}$ is $\mathcal{P}$-satisfactory (or $\mathcal{P}$-nonwasteful) on E. Without additional conditions the answer would always be affirmative. One could take the identity mappings for each μ^i, taking M equal to E, and make h equal to $\mathcal{P}$, or a suitable selection from it if $h \circ \mu$ is required to be single-valued. Such a mechanism provides enough "channel capacity" to permit each agent to communicate his entire characteristic to the others. Then any or all of them could calculate $\mathcal{P}$-optimal outcomes.

One of the criticisms of such a mechanism, which is in an obvious sense "centralized," is that it is either infeasible for every agent to communicate fully his characteristic to a center and have the center calculate the outcome, or the resources used in communication and computation would be so large as to leave too little for direct economic use.

Such considerations lead to imposing a limi-

tation of "channel capacity" through restricting the information-carrying capacity of the messages used by the process. A restricted message space M would allow some information to pass but not necessarily all information. Such a restriction would act analogously to a limitation of the (cross-sectional) diameter of a pipe restricting the flow of fluid through that pipe. When the message variables take real values, a natural restriction is to limit the number of variables whose values can be communicated, i.e., to limit the dimension of the (Euclidean) message space. However, a technical difficulty arises due to the fact that it is possible to "smuggle" two variables by encoding them in the value of one variable and then recovering the two values at the other end. The same phenomenon exists even when the messages are allowed to have a more general qualitative nature than the values of real variables. "Smoothness" or regularity conditions must be imposed on the communication process in order to make restrictions of information carrying capacity meaningful. Two types of conditions have been given: one by Hurwicz (1972a) applies to the case when the message space is Euclidean, the other given by Mount and Reiter (1974a, 1974b) applies to topological message spaces.[5]

When such conditions are imposed on a mechanism its performance is thereby restricted. The question arises whether there are any mechanisms which meet the conditions and if so, for what class of economies are there Pareto satisfactory mechanisms of this type. This problem is analogous to the problem of character-

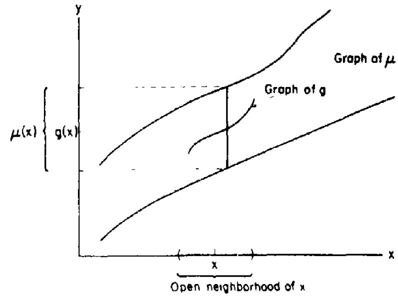


FIGURE 2

izing the subset of classical economies for which competitive equilibrium exists. Mount and Reiter (1975) have studied this problem. Their results are summarized next.

Using the definition of a process which makes the composition $h \circ \mu$ single-valued, they consider a collection of such processes, which they call a *mechanism*, all using the same message space and outcome function, and show that a mechanism whose message correspondences are locally threaded can be $\mathcal{P}$-satisfactory on a class of environments E if and only if the correspondence $\mathcal{P}$ is a union of continuous functions (a completely threaded correspondence) on E.

These theorems identify a property (complete threading) which when applied to the Pareto correspondence gives a condition which is necessary and sufficient for the existence of a (decentralized) mechanism whose performance is Pareto-satisfactory. They then ask the question, On what class of environments does the Pareto correspondence have that property? I.e., on what class of environments is the Pareto correspondence completely threaded? Equivalently, on what class of environments can all Pareto optima be achieved by decentralized means? They show that, in the presence of certain rather standard conditions on economies, if preferences are strictly monotone, then the Pareto-utility frontier correspondence (associating with each economy its Pareto frontier in the space of utility values of the participants) is completely threaded. Examples show that strict mono-

[5]The Mount-Reiter condition requires that the message correspondence be *locally threaded*. A correspondence (from one topological space to another) is locally threaded if at every point of its domain there is an open neighborhood of the point on which a continuous function is defined whose graph is inside the graph of the correspondence. (By the graph of correspondence F with domain X and range Y is meant the subset of pairs of points (x, y) in $X \times Y$ such that y is in $F(x)$.) In the case of a correspondence from the real line to the real line a typical picture is as shown in Figure 2. The term "locally threaded" comes from the fact that the graph of the function g runs through the graph of the correspondence μ like a piece of thread.

tonicity is indispensible. If, in addition, the set of points (allocations or trades) which are Pareto equivalent to a given Pareto optimal point is a singleton (Pointedness Assumption), then the "Contract Curve" correspondence is also completely threaded. (The contract curve correspondence associates to each economy the set of its Pareto optimal allocations.)

The classical welfare theorems establish the Pareto-satisfactoriness of the competitive mechanism on the class of convex environments. However, it is not known whether the competitive mechanism has a locally threaded message correspondence on the full class of environments on which the welfare theorems hold. It was established by Mount and Reiter (1974) that the competitive mechanism does satisfy that regularity condition on the class of pure trade environments with Cobb-Douglas utilities. Furthermore, it is clear that when the competitive equilibrium is unique (and the Walras correspondence is upper hemi-continuous), the regularity condition (local threading) is also met. The case of multiple equilibria for environments which do not satisfy the assumptions of their theorems remains open. It should be pointed out that the competitive mechanism, as it is ordinarily specified, does not meet their requirements for a mechanism when it is applied to economies in which it has multiple equilibria. Mount and Reiter in effect require that a particular equilibrium be selected in a continuous fashion as the environment varies. Indeed, one interpretation of their results that the conjunction of (i) the requirement that the economic mechanism make a selection of equilibria for cases in which there are multiple equilibria, (ii) the regularity condition on communication, and (iii) Pareto-satisfactoriness of performance restricts the allowable environments to very classical ones.

Another class of questions posed by the (new)[2] welfare economists relates to the communication capacity (informational size of the message space) needed in order to have a mechanism that achieves specified performance on a given class of economies. Questions of this type were posed by Hurwicz (1972b) for mechanisms that are restricted to Euclidean message spaces. He considered the classical economies, for which the competitive mechanism is nonwasteful, and asked whether there is any other mechanism whose performance is nonwasteful on the same class of economies, but whose message space is of lower dimension than that of the competitive process. He showed that there can be no such process. (To exclude "smuggling" of information, Hurwicz uses the condition that the lower inverse of the message correspondence be *quasi-Lipschitzian*, i.e., have a selection which satisfies a uniform Lipschitz condition.) (See e.g., T. M. Apostol, 1957.)

Mount and Reiter (1974), using their concept of informational size of message spaces and the (regularity) condition that the message correspondence be locally threaded, showed independently that there can be no other mechanism whose performance is the same as that of the competitive mechanism (e.g., whose outcomes are competitive equilibria) which uses a message space[6] locally informationally smaller than that of the competitive mechanism.[7]

M. Walker, exploring the concept of informational size introduced by Mount and Reiter, gave definitions of some stronger related concepts. Hiroaki Osana, using one of these extended concepts of informational size and using the Mount-Reiter regularity condition (locally threaded message correspondence), showed that there can be no process which is nonwasteful in classical economies and whose message space is informationally smaller, in the modified sense, than that of the competitive process.

These theorems establish the minimality (in an appropriate sense) of the competitive process in terms of the channel capacity required to

[6]Message spaces are required to be Hausdorff spaces. In a Hausdorff space points may be separated from one another by a pair of nonoverlapping open neighborhoods, each containing one of the points.

[7]A related result is the characterization of the competitive mechanisms given by Hugo Sonnenschein.

achieve nonwasteful performance in classical economies.

The same questions are raised about mechanisms designed to work in nonclassical (nonconvex) economies. Are there informationally decentralized mechanisms whose performance is Pareto satisfactory in nonclassical economies? What are the informationally requirements, in terms of size of message space of mechanisms designed to perform nonwastefully in nonclassical economies?

A limitation on the size of this message prevents one from reporting this work in detail here. Mechanisms designed for nonclassical economies have been studied by Hurwicz (1972a, 1972b), where the focus is on externalities, Hurwicz (1960), Hurwicz, Roy Radner and Reiter (1975), Hidei Kanemitsu, Arrow and Hurwicz, Xavier Calsamiglia, and also by Masahiko Aoki (1970b), Antonio Camacho, J. H. Dreze and D. De la Vallee Poussin, Geoffrey Heal (1960), J. O. Ledyard (1968, 1971), John S. Chipman and Aoki (1967). The general import of this work is that the presence of nonclassical properties such as externalities, indivisibilities or increasing returns increases the informational requirements of mechanisms; usually a finite dimensional message space does not suffice in such cases.

Study of the informational properties of mechanisms has been mainly concerned with communication. However the internal computations performed by agents, including the center if there is one, are also important. While several interesting attempts have been made to compare computational requirements of different mechanisms (Çarl Futia, Hajime Oniki) and in other ways to take account of the effects of limited computational capacity on the performance of mechanisms (Roy Radner 1973, 1975a, Radner and Michael Rothschild 1975 and Radner 1975b), this area remains open to exploration and is of great importance for our better understanding of economic systems.

REFERENCES

Masahiko Aoki, "Increasing Returns to Scale and Market Mechanisms," Technical Report No. 6, Institute for Mathematical Studies in the Social Sciences, Stanford University 1967.

———, "Two Planning Algorithms for an Economy with Public Goods," Discussion Paper No. 029, Kyoto Institute for Economic Research, Kyoto University 1970b.

T. M. Apostol, *Mathematical Analysis,* Reading 1957.

Kenneth J. Arrow, "An Extension of the Basic Theorems of Classical Welfare Economics," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* J. Neyman, ed., Berkeley 1951, 507–32.

——— **and Gerard Debreu,** "Existence of an Equilibrium for a Competitive Economy," *Econometrica,* 22, 265–90

——— **and Leonid Hurwicz,** "Decentralization and Computation in Resource Allocation," in R. W. Pfouts, ed., *Essays in Economics and Econometrics,* Chapel Hill 1960, 34–104.

Xavier Calsamiglia, "On the Possibility of Informational Decentralization in Non-Convex Environments," Ph.D. Dissertation, University of Minnesota.

Antonio Camacho, "Externalities, Optimality and Informationally Decentralized Resource Allocation Processes," *Intern. Econ. Rev.,* 1970, *11,* 318–27

John S. Chipman, "External Economies of Scale and Competitive Equilibrium," *Quart. J. Econ.,* Aug. 1970, *84.*

Gerard Debreu, *Theory of Value,* New York 1959

H. D. Dickinson, "Price Formation in a Socialist Community," *Econ. J.,* Dec. 1933, *43.*

J. H. Drezé and D. De La Valleé Poussin, "A Tatonnement Process for Guiding and Financing an Efficient Production of Public

Goods," Discussion Paper No. 6922, Catholic University of Louvain 1969.

Carl Futia, "The Complexity of Economic Decision Rules," Part I and II, Technical Report OW-5, Center for Research in Management Science, University of California, Berkeley 1974.

A. Gibbard, "Manipulation of Voting Schemes: A General Result," *Econometrica,* 1973, *41,* 587–601.

Theodore Groves and John Ledyard, "An Incentive Mechanism for Efficient Resource Allocation in General Equilibrium with Public Goods," Discussion Paper No. 119, Center for Mathematical Studies in Economics and Management Science, Northwestern University 1974.

——— **and** ———, "The Existence of an Equilibrium under an Optimal Public Goods Allocation Mechanism," manuscript 1975.

R. Guesnerie, "Pareto Optimality in Non-Convex Economies," *Econometrica,* 1975, *43,* 1–29.

Friedrich A. von Hayek, "The Present State of the Debate," in *Collectivist Economic Planning,* London 1935, 201–43.

———, "The Use of Knowledge in Society," *Amer. Econ. Rev.,* 1945, *35,* 519–30

Geoffrey M. Heal, "Planning, Prices and Increasing Returns," *Rev. Econ. Stud.,* 1971, *38,* 281–94.

———, *The Theory of Economic Planning,* Amsterdam 1973.

Harold Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica,* 1938, *6,* 242–69.

Leonid Hurwicz, "Optimality and Informational Efficiency in Resource Allocation Processes," in *Mathematical Methods in the Social Sciences,* 1959, 27–46.

———, "On Decentralizability in the Presence of Externalities," paper presented at the San Francisco meeting of the Econometric Society, 1966.

———, "On the Concept and Possibility of Informational Decentralization," *Amer. Econ. Rev. Proc.,* May 1969, *59,* 513–54.

———, "On Informationally Decentralized Systems," in C. B. McGuire and R. Radner, eds., *Decision and Organization,* Amsterdam 1972a, 1–29.

———, "On the Dimensional Requirements of Informationally Decentralized Pareto-Satisfactory Processes," presented at the Conference Seminar in Decentralization, Northwestern University Feb. 1972.

———, "On the Existence of Allocation Systems Whose Manipulative Nash Equilibria are Pareto-Optimal," presented at 3rd World Congress of the Econometric Society, Toronto, Aug. 1975.

———, "On Informational Requirements for Non-Wasteful Resource Allocation Systems," unpublished, April 28, 1976.

———, **Roy Radner and Stanley Reiter,** "A Stochastic Decentralized Resource Allocation Process: Part I," *Econometrica,* 1975, *43,* 187–221; "Part II," 363–93.

Hidei Kanemitsu, "On the Stability of an Adjustment Process in Non-Convex Environments," presented at the Second World Congress of the Econometric Society, England, Sept. 1970.

Tjalling C. Koopmans, *Three Essays on the State of Economic Science,* New York 1957.

John O. Ledyard, "Resource Allocation in Unselfish Environments," *Amer. Econ. Rev. Proc.,* May 1968, *58,* 227–37.

———, "A Convergent Pareto-Satisfactory Non-Tatonnement Adjustment Process for a Class of Unselfish Exchange Environments," *Econometrica,* 1971, *39,* 467–99.

Abba P. Lerner, "Statics and Dynamics in Socialist Economics," *Econometrica,* 1937, *43,* 1–29.

K. Mount and Stanley Reiter, "The Informational Size of Message Spaces," *J. Econ. Theory,* 1974a, *8,* 161–91.

——— **and** ———, "Economic Environments

for Which There Are Pareto Satisfactory Mechanisms," Discussion Paper No. 124, Center for Mathematical Studies in Economics and Management Science, Northwestern University, *Econometrica*, forthcoming.

Hajime Oniki, "Theoretical Study on the Cost of Decision-Making Economic Organization," 1973.

Hiroaki Osana, "On the Informational Size of Message Spaces for Resource Allocation Processes," presented Apr. 1976 at the National Bureau of Economic Research Conference Seminar on Decentralization, Northwestern University.

Roy Radner, "Aspiration, Bounded Rationality and Control," Presidential Address, Oslo meeting of the Econometric Society, 1973.

———, "A Behavioral Model of Cost Reduction," *Bell J. Econ.*, 1975a, *6*, 196–215.

——— **and Michael Rothschild,** "On the Allocation of Effort," *J. Econ. Theory*, 1975, *10*, 358–76.

———, "Satisficing," *J. Math. Econ.*, 1975b, *2*, 253–62.

Stanley Reiter, "The Knowledge Revealed by an Allocation Process and the Informational Size of the Message Space," *J. Econ. Theory*, 1974a, *8*, 389–96.

———, "Informational Efficiency of Iterative Processes and the Size of the Message Space," *J. Econ. Theory*, 1974b, *8*, 193–205.

Paul Samuelson, "The Pure Theory of Public Expenditures," *Rev. Econ. Statist.*, 1954, *36*, 387–89.

Hugo Sonnenschein, "An Axiomatic Characterization of the Price Mechanism," *Econometrica*, 1974, *42*, 425–34.

M. Walker, "On the Informational Size of Message Spaces," Economic Research Bureau, Working Paper No. 149, State University of New York, Stony Brook 1975.

Marginal Cost Pricing in the 1930's

By ABBA P. LERNER*

I have to treat this assignment as a chapter for my autobiography, since I cannot separate it from the beginning of my interest in economics.

My first contact with economics was the introductory lecture of a course in economics in evening classes offered by the London County Council in the East End of London in 1919. The lecturer was a young man from Cambridge who spent the whole hour telling us how wonderful was a certain Professor Alfred Marshall. I did not believe anyone could be as wonderful as all that and I did not continue with the course or have any further contact with economics until I went to the London School of Economics exactly ten years later

I came to the London School of Economics *(LSE)* as a socialist, with Marxist inclinations and with some rather grandiose notions of turning bourgeois economics to socialist use. I could not yet have been cured of my first vision of the Russian Revolution as the emancipation of man from subservience to money, the root of all evil, since I remember being utterly confused when Professor Postan showed me a letter from a Russian professor who had been sent to Siberia for bourgeoise deviations in his writings in ichthyology. I could not believe that Postan was lying, or that he had been deceived. On the other hand I was not ready to believe that Russia had become such an insane tyranny, so that for quite a while this remained an "unsolved puzzle."

I fell in with a group of revolutionary socialists at the school, finding great enlightenment in the idea of The Class Struggle, until I decided that although explaining the world in terms of two classes was an improvement on supposing perfect harmony and community of interests, a closer approximation to the real world was obtainable by considering more than two interest groups. In this escape from the Marxian duality I was helped by observing in a fellow Marxist, a convert from some Christian fundamentalism, a horror at something I had said which he saw as an incredibly heretical departure from "belief in" the class struggle. But my full emancipation may have come much later. Indeed I remember meeting Max Radin in Berkeley in 1938 when he teased me on having flirted with the Fourth International instead of sticking with the democratic Second International, and I provided him with a surprisingly durable source of amusement by telling him that it was because I had learned to count beyond two.

My own shift of interest from the Third to the Fourth International, from Communism to Trotzkyism, occurred in 1931. (I say shift of interest rather than conversion, because, although there was a period in which the class struggle seemed to me a useful abstraction, I was never able to accept either the labor theory of value or dialectical materialism.) In 1931, a communist fellow student, who was proficient in Russian, had come back from a tour of the Ukraine as a translator and reported the impending mass starvation imposed on the peasants by the forced grain collections. For this he was denounced as a Trotzkyist and he then indeed became one, and I became a Trotzkyist "fellow traveller."

Before I began my studies at *LSE,* a friend who had been there for a year or two, got me to read a couple of books to prepare for *LSE.* I remember being quite impressed with the idea of consumer's surplus, and that he tried, unsuccessfully, to persuade me that it was not a good idea at all. It was apparently looked down upon at *LSE.* I had also read Henry George, and found him quite impressive, both on free trade and on "the single tax" on land values, and had been completely taken in by Veblen as the only man who dared bring scientific method to social

*Queens College, City University of New York

problems. I was thus greatly puzzled by the amusement displayed by John R. Hicks, who was my tutor in the first year, when I told him what I had read in economics.

At the end of my first year I sat down to consider what I had learned during the year. While my wife was complaining that all she ever heard when I was talking with my friends was "equilibrium," it seemed to me that the only thing I had learned was the idea of "marginal." I was quite depressed. It seemed to me that I should have learned that in a few hours at the most But since then I have come to wish that more of my students would be able to master the concept of "marginal" properly by the time they sat for their bachelor's degree.

This brings me to the topic of my paper. I became a rabid marginalist, fanatically enthusiastic about the principle that economic efficiency, which is, of course, socially desirable, required every product price to be equal to the marginal cost (more strictly to the value of the additional factors required to produce an additional unit of the product where this is equal to the value of the alternative products sacrificed)

I remember reaching the conclusion that retail stores should charge the marginal cost of making each sale, and since it took the salesman no more time to sell a larger quantity than a smaller quantity, the addition to the cost of the material should be the same, so that the total price would be less per unit the greater the amount bought. I toyed with the idea of approximating this by having the shops sell everything at cost, and charge the customer a fee for entering the shop, which would pay for the provision of the retailing services, and with similar extensions of the marginal cost pricing principle. My fellow socialists were outraged at such devices, which would penalize the poor, and I was forced prematurely to the conviction that what was wrong with the poor was not the prices they have to pay but that they had too little money—that the solution of poverty lay not with the manipulation of prices but with the distribution of money income.

My enthusiasm for marginalism did not at first have any special connection with socialism or with socialist pricing. It first was directed to the social inefficiency of monopoly in raising price above marginal cost and thus interfering with my newly discovered Invisible Hand and its promise of economic efficiency. I had many arguments with those who, in objecting to the monopolist charging more than the *average* cost, seemed to me to be carried away by anger at (or perhaps jealousy of) the profits enjoyed by rich monopolists, or by deep concern at the exploitation of the poor customers who had to pay his exorbitant prices. To me these seemed irrelevant and sentimental interferences with the economist's job of avoiding the waste, the misallocation of resources, from prices above or below the marginal cost. Anger or jealousy about the receiver of the "monopoly profits" and concern about the exploitation of customers who had to pay the monopoly prices, obscured the economics. They distracted attention from efficient versus inefficient use of resources, to the quite different problem of some people being richer and some poorer than one would like them to be. Here too I felt that the issue was one of the distribution, or perhaps the redistribution, of *money*, and furthermore, it was not really the redistribution between monopolistic sellers and the "exploited" buyers of such goods, but between rich and poor. It was perfectly possible for "exploiting" monopolists to be poorer than "exploited" customers. Once more I was led to the conclusion that excessive wealth and excessive poverty were matters concerning not the prices of commodities but the distribution of income and wealth.

It seems obvious that socialist societies must be just as concerned with economic efficiency as capitalist societies, or at least that economists, concerned with efficiency, must see that efficiency is a matter of importance for both forms of social organization But I do not think I thought of marginal cost pricing as particularly relevant to socialism until I found myself in debate with socialists who objected to the view that socialist societies should make use of this principle.

I had an exchange of articles with Maurice Dobb who seemed to believe, at the time, that pricing of any kind was inconsistent with the rational planning of the economy in physical terms which would be possible when capitalism had been replaced by socialism, so that neither marginal cost nor average cost had any place in socialism.

Durbin and Dickinson and a number of others argued in favor of socialist societies charging the average cost. I remember especially having long discussions on this with Dickinson at a meeting of English economists in Oxford. It seemed to me (naturally) that all the logic was on my side, and that only irrelevant arguments were put forward against marginal cost pricing in a socialist society. These stressed such issues as the difficulty of figuring the marginal cost, while the average cost could easily be reached (after the event). One only had to divide what had turned out to be the total cost by the number of units sold. Sometimes they dwelt on the difficulties of distinguishing between long-run and short-run marginal cost.

I cannot here go into the details of these objections to marginal cost pricing and their refutation. I came away from these discussions with the feeling that the widespread popularity of average cost pricing had two different causes One was a failure to distinguish between the concept of perfect competition, a *means* for achieving economic efficiency that is possible under some circumstances, and economic efficiency itself as the *end* which perfect competition could serve. Price equal to average cost, one of the results of perfect competition, was mistakenly assumed to be the condition necessary for economic efficiency. The other and more important cause of addiction to average cost pricing was a happy point of coincidence of certain elements in the basic prejudices of sentimental socialists and sentimental capitalists. The socialists felt that profit was immoral and even wicked, so that they could not accept any policy that permitted price to be above average cost and thus gave rise to the objectionable phenomenon of profits. The capitalists felt that losing money must mean that something is basically wrong, since the natural purpose of economic activity is to *make* money. Furthermore they objected strongly to the subsidies that would have to be provided where marginal cost pricing would make the price less than the average cost, and to the taxes that would be necessary to pay for the subsidies. Average cost pricing thus formed a happy meeting ground in which the capitalist and the socialist sentimental prejudices could coexist with the minimum of friction.

It was at this stage that Oscar Lange came to London and wrote an article on market socialism for the *Review of Economic Studies*, of which I was the managing editor. He laid down the same principle of charging the marginal cost for everything, and I was of course delighted to have this article in the *Review*, as were all the other editors. But Lange got caught in the puzzle of what to do if the sum of the marginal costs did not add up to the total cost, which would be bound to happen unless the marginal costs, on the average, happened to be just equal to the average cost.

Lange's solution was to make a compensating adjustment to the wage, multiplying the price paid for labor by a bonus proportional to the wage, which made the pay greater than the marginal cost by just enough to absorb the surplus.

My contribution was to point out that this would be a departure from the efficiency principle. The surplus could be distributed in any way whatsoever that seemed good to the authorities, provided it was *not* related to the wage Relating it positively to the wage would not only discriminate in favor of those who already had a higher wage, aggravating the inequality of income regrettably required for economic efficiency, but would also cause a departure from the efficient use of resources. Lange quickly saw my point and accepted the correction.

Strangely enough, on rereading my article, I find it repeating the very error in Lange's article which I was correcting. I speak of the principle of having prices of products *proportional* to

marginal cost as a sufficient condition for efficiency in the allocation of resources between the production of different products. If the price of a product is, say, twice the marginal cost, all is still well if the price of the product is also twice the marginal cost in all the alternative uses of the factors. If $2 is the price of a product whose marginal cost is $1, the resources set free by producing one unit less of this product could be used to produce two dollars worth of other products (with a marginal cost of $1). One could say that the *social* marginal cost is really $2, the value of the alternative product or products that had to be sacrificed to make possible the production of two dollars worth of *this* product. The $1 is the *private* and not the *social* cost. Distributing the surplus *in proportion* to the wage, as initially proposed by Lange would still leave price *proportional* to marginal *private* cost and equal to marginal *social* cost. I was thus doing exactly what I was criticizing Lange for doing, even while successfully persuading him that he was wrong!

The trouble with Lange's original rule, as well as with the formulation in my "correction" of his rule, was that the rule cannot be applied universally in all the other uses of the factors. The use of labor time for leisure, for "do it yourself" or for private exchange of services with a friend or neighbor made the private marginal cost *equal* to the price or value of the product, so that universal *proportionality* of marginal cost to price could be attained only by the special proportionality of *equality* of price and marginal cost.

This more satisfactory formulation was not developed until later. (It is in my *Economics of Control*, published in 1944, twelve years after I had started working on it.) Nevertheless, the agreement between Lange and myself on the necessity of having the wage *equal* to and not merely *proportional* to the marginal product, shows that although my formulation was faulty our understanding of the issues was correct.

It somehow seemed natural, both to Lange and to myself, in those days, to suppose that the revenues from the sale of all products at marginal cost would exceed the total cost, i.e., that in general the marginal cost would be above the average cost. There would then be a surplus available which I later (in my *Economics of Control*) called a "social dividend" to be distributed equally among the whole population as the simplest and most efficient way of alleviating poverty or increasing equality. We did not anticipate the growth that has taken place in government spending on armaments and on much more complicated and much less efficient ways of helping the poor and various other groups disguised as deserving poor. (I have never been able to understand the scorn poured by Marxists on Bernard Shaw's insistence of equality as the essence of the socialist ideal.) More recently it has received much publicity as Milton Friedman's "negative income tax."

In my gropings toward the maximum social benefits to be derived from marginalism I had originally given practically no attention to administrative problems, and had almost automatically pictured socialist society as some sort of universal government enterprise which would instruct all the managers, who would be government employees, to follow the marginal cost pricing principle. I remember being surprised by Lange's acceptance of naturally small private enterprises, "farmers and barbers," as perfectly compatible with socialism. I had never thought of this but found it immediately acceptable. In general it seems to me that no disagreement between Lange and myself on economic theory ever survived an hour's discussion, although I understand that many economists have written on differences between our approaches of which I am quite unaware.

Perhaps marginalism has continued to be my hangup, but I still feel that this principle is not fully exploited by many modern economists. The egalitarianism behind the suggested equal division of Lange's excess of average price over average marginal wage cost, my "social dividend" or Friedman's negative income tax, I see *economically* justified as a more efficient way of using income to provide satisfaction to individuals. It is based on my assumption that other people, like myself, have their well-being in-

creased by having more income, and that, like myself, they experience diminishing marginal utility of income as a result of some element of rationality in their use of their income. They quite often choose to spend their money on what they find *more* rather than *less* satisfying, so that an additional dollar means less to a man when he is richer if he can spend it only on the less satisfying goods he avoids buying when he is poorer. In the absence of knowledge of the *absolute level* of the satisfaction of others, the best way of distributing the surplus (over what is necessary to induce the production of the income) is to divide it equally. This is how my learning some economics at *LSE* gave a rational meaning to my egalitarian impulses.

The empathy that makes me attribute utility, rationality and diminishing marginal utility, to individuals other than myself does not extend to society. I do not attribute to society either a level of utility or a diminsihing marginal utility of income. While an individual's preferences can indeed be considered as an *ordering* of social states, I see no reason for assuming that social decisions involve a social ordering. This assumption seems to me very much like the assumption that statisticians used to make that index numbers ought to behave just like natural numbers and conform to the reversal test, so that Paasche and Laspeyres would have to be identical.

Thus when Kenneth Arrow says that "the outcome of a social choice procedure (whatever you call it) is an ordering," I get lost. I would think that the outcome of a social choice procedure would be not an ordering but a *choice*—a choice of a social policy. This choice will, of course, result in social states which some individuals will look forward to with hope, and some with dread.

What I find much more important than the, to me idle, talk of social orderings is that the restriction of *individual* preferences to *orderings* constitutes a flight from marginalism. The various invariances of individual orderings with respect to monotonic transformations, only serve to banish the assumption that others, like myself, have diminishing marginal utility of income, and destroys the rational basis for egalitarianism as efficiency.

But egalitarian impulses will out. Other rationalizations are sought, and this has led to a plethora of other, or apparently other, reasons for advocating the egalitarian distribution of surplus. Rawls finds it in fairness and/or justice. His "original state" allegory is certainly a most dramatic way of saying "let us not be biased," but in his "maximin" principle I see no basis (in spite of his denials) other than the assumption of diminishing marginal utility of income or an infinite risk aversion, or perhaps both.

Arrow and Amartya Sen find it in "*equity*." I have just had some intensive discussion on this subject with Sen. He rejects the utilitarianism on which I base my case for the equal distribution of surplus. Instead he turns to equity (or, more modestly, to his "weak equity principle") aimed not at maximizing utility but at *equalizing* it. He would always, therefore, advocate the transfer of "some" income from happier to less happy individuals, irrespective of their income. The "some," I think, betrays Sen's discomfort at having to take income from a happy poor person and give it to a miserable rich person, even if the loss to the former is greater than the gain to the latter. (You would presumably know if this was the case if you knew which was the happier and which the more miserable.) I see neither the possibility nor any desirability in equalizing the utility levels of different individuals, nor indeed of any or "some" movements in that direction.

Arrow's equity principle seem to come to the same thing as Sen's. I see in both a retreat from marginalism, reminiscent of the addiction of pre-Lange-Lerner socialist egalitarians to the principle of having price equated to average cost, but, of course, on a much more sophisticated level. They are forced by their abhorrence of the subjective basis for assuming individual diminishing marginal utility of income, *without comparing them*, to rest their egalitarianism on the much less defensible *comparison* of the actual levels of individual total utility.

ABRAM BERGSON, Harvard University: Kenneth Arrow expatiates on the theory of social choice, a discipline that he himself has done so much to create and shape. He apparently continues to hold a view of the relation of that discipline to welfare economics which he took when he first expounded on social choice theory in his now celebrated *Social Choice and Individual Value*. I refer to his identification of the social ordering function of social choice theory with the so-called social welfare function of welfare economics. Despite criticism by diverse writers, including I. M. D. Little, Paul A. Samuelson and me, such a view of the two functions seems to have been espoused through the years not only by Arrow but by others as well.

True, in response to the criticism, Arrow has agreed that it might be better to refer to the function on which he focuses in social choice theory as a "constitution," rather than as a social welfare function. In his paper here, he indicates that that is now his preferred usage. But in adopting that usage, he has also affirmed that the "difference" between the two functions "is largely terminological." The two functions obviously have a certain kinship to one another. I for one am happy to acknowledge that kinship. But the intended burden of criticism of Arrow's identification of the two functions has from the beginning clearly been that the difference between them is much more than terminological; that there is in fact an important substantive difference as well. The difference is viewed as important for the reason that Arrow's Impossibility Theorem, while applying to the social ordering function of social choice theory, is seen as having at most only a very tenuous relation to the social welfare function. Not too surprisingly, therefore, Arrow's verbal concession has not really quieted dissent. In effect, then, the methodological issue that was first posed by *Social Choice and Individual Values* a quarter of a century ago still remains unresolved.

That issue is also one that I can hardly resolve here, but I should record that I am among those who, despite Arrow's verbal concession, continue to have misgivings about his view on the two functions. Perhaps I should explain also that in a forthcoming article (in the *Journal of Public Economics*) I try again to explain my standpoint. How convincingly I do so remains to be seen, but let us hope that the fundamental matter in question will not be allowed to remain controversial much longer.

THOMAS MARSCHAK, University of California, Berkeley: Stanley Reiter's survey skillfully strips to the essentials a variety of achievements in a new, important, and difficult field. No longer can the field be viewed as exotic and inaccessible. Now, equipped with this survey and with Leonid Hurwicz's Ely Lecture of four years ago, the interested newcomer who wants to learn what is known has a useful chart of what lies before him, and finds cogent reasons for starting the journey

Still the field, unlike the "old" new welfare economics, is not quite ready for the textbooks. It seems to have struggled out of its infancy, but at best it is a promising though quite unsteady toddler. The reason, I believe, is that several dark but crucial corners have yet to have much light shone into them. I want to comment briefly on three of these dark, mysterious corners and their relation to Reiter's survey; they do not directly appear in the survey—quite properly, I think, since it is a survey of achievements and not of puzzles.

The three dark corners are *first*, what I shall call Performance Theorems as opposed to Possibility Theorems. *Second*, the modeling of the information technology needed to run allocation mechanisms. And *third*, what I shall call Behavioral Theories of allocation mechanisms as opposed to Robot Theories.

First, most of the work which Reiter surveys asks: "Can mechanisms with certain properties exist for economies with certain properties?" For the case of the competitive mechanism, one main group of results says that what the competitive mechanism achieves *at equilibrium* in the economies where it works well, no other

mechanism can achieve more cheaply, where "cheapness" has to do with the mechanism's communication burden. These results are an important and natural extension of the "old" new welfare economics. For the "old" new welfare economics assures us that the competitive mechanism sustains every Pareto optimum in classic economies and so makes the competitive fable the standard against which real economies and real mechanisms are to be judged. The "new" new welfare economics strengthens the fable further by showing that the competitive mechanism is, in a sense, a best mechanism with regard to communication effort. At the same time it may have weakened the fable, since it reveals that the competitive mechanism may suffer from a basic incompatibility with individual incentives. But one can ask questions about allocation mechanisms which are not questions of possibility. One can ask instead about a mechanism's performance over time as the economic environment changes. That requires some measure of performance at any instant of time—some social welfare function in the original Abram Bergson sense, defined on the current actions of the economy's agents and on the current economic environment. One wants to study the sequence of values of the performance measure which a mechanism achieves as it generates actions in response to the sequence of environments. Ideally, the performance measure would be a "net" measure, that is, it would subtract the resources consumed in operating the mechanism itself. From this point of view, the ultimate question about the competitive mechanism is not "are cheaper mechanisms which achieve exactly the same thing at equilibrium possible?" but rather "are mechanisms achieving a better net performance over time possible?" It is certainly not excluded, for example, that a mechanism which foregoes optimality at equilibrium might yet be a better mechanism in the sense of net performance.

This type of question is, of course, an extremely ambitious one to study in models of a full-scale economy, whereas the possibility questions surveyed by Reiter are not, and are therefore a natural starting place. But if one turns to "miniature" economies, then performance questions are not beyond hope. By a miniature economy I mean, for example, a firm with branch plants, each in the charge of a manager who observes part of the environment, receives from others whatever message the chosen mechanism gives him, and chooses local production actions as the mechanism specifies. The firm's current profit measures its current gross performance, and current net performance is obtained by subtracting suitable amounts for the mechanism's operating costs. The work of Theodore Groves and Roy Radner, for example, is in this spirit and compares several mechanisms for a firm (team) with a quadratic gross performance measure.

Second, the information technology required for running mechanisms will need to be modeled well if we are ever to be more than casual and arbitrary in assessing a mechanism's costs. Reiter deals with smoothness conditions which various studies impose on the communication implied by a mechanism—the way in which messages and actions change as the environment changes must be locally threaded, or Lipschitzian, depending on the possibility result to be established. He motivates such conditions by pointing out that without them allocation mechanisms achieving optimality, possibly with very "small" messages, can be trivially designed. One can, for example, smuggle many numbers into one number.

What is wrong with doing so? The answer is that one would expect something like the smoothness conditions to be desirable in models of an orderly transmission technology. If, for example, a message to be communicated can be any real number in a certain interval, then a communicator who uses a transmission line would have to approximate the message and code it into a finite number of symbols, that is to say, a *grid* would have to be imposed on the interval. The finer the grid, presumably, the more costly. It would generally be undesirable if small changes in the uncoded messages to be sent led to large changes in the messages recognized after coding, transmission, and decoding. For then the responder to the message may often take inappropriate actions. Fineness of the grid—closeness to smoothness, in some ap-

propriate sense—makes such jumps less frequent. Ultimately one would want to replace smoothness conditions by a balancing of the costs and benefits of fineness of grid in a carefully modeled transmission technology. But transmission is not the only task required in running a mechanism. There are computing, observing, remembering, and action-taking as well, and models of these tasks are need too. One possible approach, inspired by the theory of finite-state machines, lets everything be finite: the message language, the possible contents of a memory, the possible environmental observations to be distinguished, the possible actions to be taken. Every task in a mechanism then assigns an element of one such finite set to given elements of other such sets. One then boldly says that it is the size of these finite sets which determine the mechanism's costs. Whether this approach, or some other approach, is a good one surely has to rest eventually on empirical study of the costs of running mechanisms in full-scale economies or in miniature ones.

This brings me to the *third* and last dark corner. It seems fair, now that the field is out of its infancy, to speculate as to what studies of real economic organizations the theory might some day assist. Can there ever be policy choices for designers of future real mechanisms, or empirical studies of existing real mechanisms, which the theory might help to guide? Here one problem is that a large part of the theory is, in effect, a theory of robots. In the real world one might indeed imagine a mechanism to be chosen by a super-designer, but the people who are supposed to carry it out may not do so. I conjecture that in real situations it will prove useful to view a designer as safely able to choose not a whole mechanism but rather four things: the person-hours (and machine-hours) to be made available for the various informational tasks, the structure of individual rewards (who gets what when the organization chooses a certain action and the environment takes a certain value), the language in which messages are written, and the set of actions from which a given person is to choose.

The designer's choice of these four objects constrains the mechanism actually used somewhat, but it is a behavioral question as to which of the possible mechanisms finally comes into use: the organization's members will, in effect, make the choice. Will the theory say something useful about what values of the designer's four chosables are good ones?

One might predict that the mechanism which behaviorally survives will be an incentive-compatible one in the sense of the theory. If so, that is useful for the designer to know. But the prediction rests on a view of the organization's members as rational players of a game who are persuaded by the Nash property—a daring behavioral conjecture. Still it is certainly a start.

Serious exploration of the three corners lies ahead of us. I am persuaded that on the whole good judgment has been used in giving priority to the easier corners. That good judgment has been a main reason for the successful emergence of the field from its infancy.

JERRY S. KELLY, Syracuse University: Kenneth Arrow is to be congratulated for his excellent introduction to the valuable ideas of S. L. Strasnick and Peter Hammond. I shall discuss a choice made in his exposition and show an alternate route that leads to some interesting insights into the Strasnick-Hammond results.

Arrow's technical device for exposition is to combine the $U_i(x)$ utility functions over social states to a single function, $U(x, i)$ which serves as the argument for a constitution, $f(U(x, i))$. Alternative versions of the aggregation problem are then presented by alternative invariance rules f is required to satisfy.

The other possible route is to focus on the domain elements, (x, i), of U. Such a pair, consisting of an individual and a social state, will be called an *alternative* and interpreted as "being individual i in state x." Individuals will be assumed to have "extended" preferences over these alternatives, so that for individual k to believe Arrow's "individual i in state x is better off than individual j in state y," we write

$$(x, i)\ P_k\ (y, j).$$

These extended preferences are then to be aggregated to an ordering over social states.

Within this framework, Amartya Sen follows P. Suppes in applying a grading principle of justice to the aggregation procedure. Suppose a world with only A(rrow) and H(urwicz). Then A clearly sees state x superior to y if

$$(x, A)\, P_A\, (y, A)$$

and

$$(x, H)\, P_A\, (y, H).$$

But the grading principle also has A judging x to be superior to y if

$$(x, A)\, P_A\, (y, H)$$

and

$$(x, H)\, P_A\, (y, H),$$

capturing some of the Rawlsian "veil of ignorance" that keeps A from using the information about who he will be in each state. The grading principle of justice is then applied to aggregation by having x socially preferred to y if everyone finds x to be superior to y in the above manner. Sen, however, found that with all possible profiles of extended preferences allowed, the use of the grading principle in aggregation leads to contradictions. He also proved that these contradictions could be avoided if you assume the Axiom of Complete Identity which requires that all individuals have identical extended preferences. Then, if this common ordering on (x, i) alternatives is representable by a utility function we are back with Arrow's $U(x, i)$ and an f satisfying co-ordinal invariance. The Hammond-Strasnick results thus reveal the existence of an aggregation procedure satisfying the grading principle of justice and suitable interpretations of the usual Arrow conditions *as long as the Axiom of Complete Identity holds*.

But this last requirement is quite implausible, demanding a very unlikely consensus. Moreover, an important result of Sen shows this requirement to be unnecessarily strong even for avoiding the contradictions of the grading principle. It is sufficient for this avoidance that each profile satisfy just the Axiom of Identity which requires

$$(x, i)P_j(y, i) \text{ if and only if } (x, i)P_i(y, i).$$

Thus if $(x, A)P_A(y, A)$ we must have $(x, A)P_H(y, A)$. But the P_A and P_H orderings may still be very different. Hurwicz may be appalled at being Arrow and have both (x, A) and (y, A) near the bottom of his extended preference ordering while Arrow, comfortable with himself, may have both of these near the top. The Axiom of Identity can be defended as being an important part of the exercise of "putting one's self in someone else's shoes."

What can be said about the possibility of satisfactorily aggregating all profiles of extended preferences satisfying this Axiom of Identity? First, the Strasnick-Hammond results no longer work as there is nothing to require consensus on who is worst off in any given social state. Arrow recognizes this when he says "it is essential to the present construction that the comparisons of individual i in state x with individual j in state y be the same whether the comparison is made by i, j, or a third individual, k." But our new problem cannot even be expressed in Arrow's framework of $f(U(x, i))$ plus an invariance rule.

I have shown elsewhere that a suitable formulation of this new problem requires a slight variation of the notion of decisiveness (used, for example, in expressing the requirement of nondictatorship) to avoid conflict with the grading principle. With this slight revision, I conjecture there is *no* aggregation procedure satisfying the grading principle and suitable restatements of the Arrow conditions for a domain consisting of all profiles satisfying the Axiom of Identity. Indeed, using one additional revision of decisiveness, an impossibility result of this sort has been proven elsewhere. When we abandon the implausible Axiom of Complete Identity, we not only give up the Strasnick-Hammond solution, we abandon all possibility of reasonable solution. The hopes that Strasnick and Hammond raise for the use of the extra information in extended preferences for making good social choices are doomed in realistic domains even if we abstract away from the obvious problems of cheaply collecting true extended preferences.

EQUILIBRIUM IN MARKETS WHERE PRICE EXCEEDS COST

Uncertainty, Production Lags, and Pricing

By DENNIS W. CARLTON*

Availability is an important attribute of a good in many markets. Such markets include retail stores, restaurants, hotels, manufacturing, taxi cabs, airlines, parks and public utilities. Some of these markets are competitive, some noncompetitive, and some regulated. Fluctuating delivery time can be thought of as the consumers' equivalent to varying availability of a good. Here too, customers face some risk of being unable to obtain goods when they want them.

There are good reasons why some markets do not always clear in the classical supply and demand sense at each instant. The three features that characterize these markets are temporary price inflexibility, demand uncertainty, and production lags. Prices do not instantaneously adjust in response to shifts in demand. If demand is especially heavy during the morning, prices do not rise in the afternoon. Several justifications for such temporary price inflexibility are possible. Consumers may dislike price fluctuations, and firms may be providing a service of stabilizing prices in the very short run. Changing price may be costly. To provide an effective "signal," prices may have to remain fixed for some time period. Unless it is evident that demand and supply have permanently shifted, firms may be reluctant to change price. Whatever the reason, it is a fact that for many markets, price once set does not vary for some time. Of course, there still remains the issue of how the price is initially determined.

The second feature of these markets is that demand is uncertain. If demand were perfectly predictable, there would be no need to have unsatisfied customers. The final feature is that production takes time. If instantaneous production were possible, once again there need be no unsatisfied customers. We assume that recontracting or insurance markets do not develop. Such markets rarely develop in reality presumably because of high transaction and monitoring costs.

This paper discusses the implications of markets characterized by price inflexibility, demand uncertainty over the time period for which prices are inflexible, and noninstantaneous production. A competitive equilibrium is defined and its properties examined. The social welfare implications of these markets and the socially optimal policy and its relation to regulation are analyzed. This social welfare problem is exactly the same as a peak load pricing problem under uncertainty. We next deal with the behavior of a monopolist, who tends to oversupply availability, but whose behavior is consistent with a smoothly functioning economy. Finally, the issue of firm interaction and incentives for vertical integration is addressed.

I. Competitive Market Behavior

When availability is important to the consumer, the probability $1-\lambda$ of obtaining the good becomes an attribute of the good, together with its price p. Consumers maximize expected utility, and thereby generate indifference curves between $1-\lambda$ and p. These indifference surfaces need not have any particular convexity properties. Initially, we assume all consumers are identical.

The way the market operates, a firm must

*University of Chicago. I thank Franklin M. Fisher and Richard Zeckhauser for helpful comments.

decide on its price and on its amount of productive capacity before random demand can be observed. For simplicity, we assume that there is a constant per unit cost of production, and that all production takes place at the beginning of each market period before demand is observed. We assume no inventory holdings. As long as inventory costs are positive, the qualitative features of market operation would be unchanged. For any given price, the more the firm sells the higher are its profits. For any given price, the more the firm produces the greater is the probability of satisfying a customer, but the greater is the risk that resources will be invested in the production of a good that will be unsold. Holding expected profits constant, the higher the probability of satisfying a customer, the higher is the price needed to compensate for the increased risk.

From past reputation, customers have information about the probability of availability and the price that each firm offers. Equivalently, each person has information about the expected utility from going to any firm. Customers randomly frequent those firms that appear to offer the highest expected utility. In my models, customers are not allowed to search at other firms if they are unable to obtain the good at the first firm they frequent. However, as long as search costs are significant, the same type of qualitative market behavior as described below will result.

Firms realize that consumers value both price and availability. Firms compete with each other by offering the best package of price and availability until expected profits are driven to zero. Competition among firms insures that in equilibrium the utility level offered by all firms will be as high as possible subject to the constraint that expected profits are nonnegative. Equilibrium is determined as the tangency between an indifference curve and the zero profit curve in $(1 - \lambda, p)$ space. (We ignore corner solutions.)

There are two particularly interesting features of this equilibrium. First, price exceeds c, the constant cost of production. This occurs because price must compensate for unused but available goods (or, equivalently, available but unused capacity). Second, the amount supplied will not in general equal the amount demanded even in an expected value sense. Preferences of consumers for risk will determine in part whether supply will be greater or less than demand.

As the number of customers per firm grows, it can be shown (Carlton 1976b) that for the case where in equilibrium customers randomly choose among the identical firms, the equilibrium price and probability of availability will approach c and 1 respectively, where c is the constant cost of production. These are of course the equilibrium values in a classical supply and demand framework. The absolute discrepancy between supply and demand will in general grow unboundedly as the number of customers per firm increases. Lower bounds on market size necessary to achieve any given level of convergence to the classical equilibrium values of price and probability can be numerically derived. These lower bounds are derived so that they apply regardless of consumer preferences. The calculated lower bounds are very large—for example, to achieve convergence to the 1 percent level, the customer per firm ratio must at least exceed 6,500. Such large lower bounds suggest that in many situations, a serious error could be made if the uncertainty aspect of the problem is ignored.

What about the dynamics of such markets? If price exceeds marginal production cost, is there an incentive for price to fall, even though negative expected profits would result? Are these markets inherently unstable? As with the simpler traditional supply and demand model, a precise and realistic mathematical theory of non-Walrasian price setting by independent agents leading to equilibrium is not available. However, as the examples mentioned earlier illustrate, we do know that such markets exist. Although a mathematical model is lacking, it is still possible to speculate on those characteristics of a market that are likely to lead to stable market operation. One key element is the relative response lags of consumers and firms to

market conditions. If, at the beginning of a market period, one firm could announce a price slightly lower than everyone else, and thereby capture the entire market for that market period, then stable market operation seems unlikely. Also, for industries with huge fixed costs, the incentive to shade price may be so great as to preclude competition in a stable competitive market. On the other hand if consumer perceptions take time and if firms can be expected to compete in acquiring reputations as the best firms, then stable operations seem possible.

Many regulated markets, such as transportation, fit the description of the markets under study here. Arguments in favor of deregulation usually focus on contrasting the desirable features of competitive equilibrium versus the stifling effects of regulation. At least as important is determining whether the competitive equilibrium could be established. Clearly, more research on the dynamics of such markets is needed.

What happens if consumers differ in their preferences for price and availability? (Over any time period, the same person may behave as several different types of consumers.) Firms might specialize in satisfying a particular type of customer Customers who greatly value prompt delivery will frequent firms which may charge high prices but which have a high probability of delivering goods on time. It is possible, though, that such specialization can simply not occur. The reason is that in a specialized equilibrium a type 1 person could be better off at a firm serving type 2 customers. Since there are risk pooling economies of scale in these markets, each person's stochastic demand characteristics influences a firm's costs; externalities abound. This raises questions of social welfare to which we now turn.

II. Social Welfare

Given that demand is stochastic, what is the price and capacity for each firm that maximizes social welfare? Can the private market be relied on to reach the social optimum? The answers depend on the form of the social welfare function, and the way in which uncertainty enters the demand curve. The problem under discussion is identical to a peak load problem under uncertainty.

In previous models in the literature (e.g., Gardner Brown and M. Bruce Johnson, Michael Visscher), expected consumer surplus is taken as the social welfare function. For this special case, it turns out that for the model under discussion the private market can achieve the social optimum. In this optimum, expected profits are zero, and price *exceeds* the constant long-run marginal cost. This result contrasts with previous results in the peak load pricing under uncertainty literature which find that the optimal price is always less than or equal to long-run marginal cost. The main difference between those models and this one is that in this model uncertainty is entering the demand curve multiplicatively (i.e., $D(p) = x(p)u$ where u is random), and that when demand exceeds supply, rationing is random. Random rationing occurs when the consumers to be serviced are chosen randomly. In previous models in the literature, either uncertainty enters the demand curve additively and/or rationing is nonrandom. Examples of nonrandom rationing schemes would be ones where those with the highest or lowest willingness to pay were served first. It is also possible to show that when uncertainty enters the demand curve multiplicatively and when those with the lowest willingness to pay are served first, then the optimum price again exceeds long-run marginal cost, but this time expected profits are positive. A complete discussion appears in Carlton (1976c).

When the availability of a good is not assured, it is not clear that expected surplus to society is the appropriate measure of social welfare. Consumers' preferences for the risk of not obtaining the good should matter. If all consumers were identical, a social planner would maximize the utility of a typical individual. The competitive market does not in general lead to the social optimum in this more general problem. For the model with multiplicative demand uncertainty and random rationing, the condi-

tions for whether subsidization or taxation are necessary depend on the marginal utility of income (Carlton 1976b). If the marginal utility of income is higher when all varieties of goods are available, then the social optimum involves subsidizing the firms who produce the good subject to shortages. Under the reverse assumption, the firms should be run at a profit in the social optimum, and the profits taxed away.

If the government can only set price, but the firms can choose their capacity or availability level, firms will compete on capacity or availability until profits are driven to zero. If the social optimum does not involve zero profits, then this regulation process will be nonoptimal. If profits are zero in the social optimum, then a correctly regulated price will enable the market to reach the social optimum. Too high a price will lead to excessive capacity, while too low a price will lead to too little capacity. The airlines would be a good example of a market where too high a price with its consequent high level of capacity may be in effect for many routes.

III. Monopoly Behavior

The behavior of a monopolist in these markets was analyzed by Edwin Mills (1959, 1962). A monopolist is concerned with maximizing expected profits. He is not specifically concerned with inconveniencing his customers except as it affects their demand for his product when it is available. If an individual's per capita demand curve is independent of the level of availability, then a monopolist will follow the same price-output policy regardless of the consumers' preferences toward risk.

Because the monopolist ignores consumers' tradeoffs between price and availability, consumers are worse off under monopoly than under competition (Carlton, 1975, Ch. 3). For most markets, we expect that the monopolist will charge a higher price and will therefore have an incentive to provide a higher level of availability than the competitive market. Monopoly, which is usually associated with restriction of output, can provide the product to consumers more frequently than competition. In such a case, the price charged is so high relative to the increased availability that consumers are worse off than they would be in the competitive equilibrium involving a lower price and lower availability.

IV. Shock Absorbing Ability

How smoothly does an economy characterized by an interrelated sequence of these uncertain markets function? Suppose that suddenly there is an increase in the uncertainty of demand. If firms feel that the structural shift in demand is temporary, they may be hesitant to alter prices for the reasons discussed earlier, and may prefer to adjust to this shift in demand solely through changes in production (or available capacity). If firms cut back on their production, then the system is subject to more frequent bottlenecks, while if firms expand production, the system will be cushioned from the effects of this increased uncertainty. The first response tends to insure the smooth functioning of the economy, while the second response tends to destabilize the economy.

To address the issue of a smoothly functioning economy, it is useful to focus attention on the case of no inventory holdings. Firms with perishable or dated goods or firms that provide capacity (like electric utilities, airlines, or telephone companies) can be thought of as firms whose unused goods or unused capacity on one day cannot be stored for the next day. For markets with firms that do hold inventory, similar qualitative results as derived below hold, but the analytics become slightly more complicated.

Since price is fixed at p, we can write demand at p as,

$$x = \bar{x} \cdot u, \text{ or}$$

$$x = \bar{x} + e, \text{ where}$$

$\bar{x}$ is mean demand at price p, and $e = \bar{x}(u-1)$ is a random component with density $f(e)$ and cumulative density $F(e)$ defined over the range $(-\bar{x}, \infty)$. The mean and median of e are assumed to be zero.

If z is capacity and c is constant per unit capacity cost, expected profits can be written

$$\pi = \int_{-\bar{x}}^{z-\bar{x}} (\bar{x} + e)p \, dF + pz \int_{z-\bar{x}}^{\infty} dF - cz.$$

The first term is expected revenue received when demand is less than capacity, the second term is expected revenue received when demand exceeds capacity, while the third term is cost.

At fixed price p, the optimal capacity for the monopolist is found by setting $\frac{\partial \pi}{\partial z} = 0$ to find

$$p \, H(z) = c,$$

where $H(z) = 1 - F(z - \bar{x})$. $H(z)$ is the probability that demand exceeds capacity z, hence $p \cdot H(z)$ is the expected revenue from adding an additional unit of capacity onto the existing capacity z. The first order condition says that capacity should be increased until the expected revenue of an additional unit equals the cost of that additional unit. We assume the optimal price-capacity combination generates nonnegative profits, and that the second order conditions for an interior maximum are satisfied.

If the price elasticity of demand is between -1 and -2, then from Samuel Karlin and Charles Carr, it is known that, for the profit-maximizing monopolist, $p > 2c$, so that $z - \bar{x} > 0$. Now replace F with the riskier (in the Michael Rothschild-Joseph Stiglitz sense) distribution G. Under the assumption that F and G have the same median and that G is riskier than F, it follows that $z_G \geq z_F$, where subscripts refer to the corresponding probability distributions. In response to increasing uncertainty, the monopolist generally increases his production when price elasticities lie between -1 and -2.

How does the competitor react to an increase in risk with fixed prices? Differentiating the zero profit condition with respect to an index of risk, it can be shown that the competitor generally decreases his production. In a rapidly changing environment with sticky prices, the competitive market structure could create bottlenecks, while the monopolistic structure could prevent them.

V. Firm Interaction and Vertical Integration

The final issue deals with how different markets interact with each other (Carlton 1976a). If a firm that produces output buys some input from another firm, then the availability of the output depends on the availability of the input. Output firms will be concerned about not having a certain supply of input. Price of the input, when purchased in a market, exceeds the constant per unit production cost in order to compensate the input firm for the average probability of being unable to sell all its inputs (or equivalently to use all its productive capacity). If an output firm produced one unit of input for itself then the output firm would be reasonably confident of being able to use that one input, since it will use its inputs before purchasing any inputs from others. Because the output firm can initially give a higher than average probability of use to its own internally produced inputs (or input production capacity), there will be a strong cost saving incentive for some vertical integration to occur. Output firms will produce input for the predictable (high probability) component of their input demand, and pass on the unpredictable (low probability) component to others. Prices in the input market will tend to rise to reflect the change in the demand uncertainties facing input firms as a result of the vertical integration. The social welfare implications and the amount of vertical integration depend on whether the vertically integrated firms sell their input to others. In many industries, transaction costs preclude this and output firms which vertically integrate produce inputs only for their own use. Under the assumption that output firms produce input only for their own use, incentives for vertical integration (either total or partial) exist, even though everyone would be better off with no vertical integration. No vertical integration is the desired state because a pooling of risks can occur on a larger scale when there is no vertical integration than when there is.

An interesting feature of markets with avail-

ability as a characteristic is that the indifference surfaces between price and availability need not be convex. With nonconvexities, firms responding to marginal incentives can wind up at the wrong equilibrium, with the incorrect production technologies. In such situations, vertical integration can enable a firm to respond to global not marginal incentives. A vertically integrated firm will be able to coordinate its operating policy at both the input and output level. Internalization can be a way to avoid an inefficiency caused by nonconvexities.

VI

Many markets have the features of temporary price inflexibility, demand uncertainty, and production lags. The performance of these markets differs in important respects from that of markets where deterministic supply and demand curves establish equilibrium. A proper understanding of these markets is a prerequisite to a better understanding of such public policy issues as regulation, peak load pricing under uncertainty, a smoothly functioning economy, and vertical integration.

REFERENCES

Gardner Brown and M. Bruce Johnson, "Public Utility Pricing and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119–28.

Dennis Carlton, "Market Behavior Under Uncertainty," unpublished Ph.D. thesis, MIT Department of Economics, Sept. 1975.

———, "Vertical Integration in Competitive Markets Under Uncertainty," Working Paper 174, MIT Department of Economics, April 1976a.

———, "Market Behavior With Demand Uncertainty and Price Inflexibility," Working Paper 179, MIT Department of Economics, June 1976b.

———, "Pricing With Stochastic Demand," mimeo, July 1976c.

Samuel Karlin and Charles Carr, "Prices and Optimal Inventory Policy," in K. Arrow, S. Karlin, and H. Scarf, eds., *Studies in Applied Probability and Management Science*, Stanford 1962.

Edwin Mills, "Uncertainty and Price Theory," *Quart J. Econ.*, Feb 1959, *73*, 116–30.

———, *Prices, Output and Inventory Policy*, New York 1962.

Michael Rothschild and Joseph Stiglitz, "Increasing Risk: I, A Definition," *J. Econ. Theory*, 1970, *2*, 225–43.

Michael Visscher, "Welfare-Maximizing Price and Output with Stochastic Demand: Comment," *Amer. Econ. Rev.*, Mar. 1973, *63*, 224–29.

Resource Extraction with Differential Information

By RICHARD J. GILBERT*

Since the subject of this session is "Equilibrium in Markets Where Prices Exceed Costs," the topic of this paper presents an immediate difficulty, namely whether land rents should not be considered costs. I would prefer that debate on this issue be postponed until after I have used my allotted time.

The motivation for this paper arose from a study of a search model of resource exploration. The study (Gilbert) showed that when no new information is revealed in the process of search, exploration can be completely characterized by a reservation cost rule. The reservation cost is a sufficient statistic for the cost of search and extraction in each area. Furthermore, if probability distributions are known and firms are risk-neutral, the sequence of extraction would be efficient. The area with the lowest reservation cost would be and should be exploited first.

This is a reassuring and not unexpected result. Although land in this example has uncertain characteristics, these uncertainties are common knowledge. Furthermore, it was assumed that the uncertainties are independent of the actions of particular individuals. These assumptions imply that for the purpose of resource allocation, unexplored land can be considered a commodity with essentially the same characteristics as known deposits.

Now let us suppose it is possible to produce information about uncertain prospects. In the exploration of petroleum resources, a substantial amount of geological and geophysical work may precede the decision to drill at a particular area. In the United States, outlays on these information generating activities have been on the order of 20 percent of drilling costs in recent years.[1] The incentive to produce information about a particular area may depend on information produced elsewhere in the economy. Thus it is not obvious that the market will provide signals that lead to an efficient allocation. The issue here is not one of "spill-over" externalities: i.e., information about one area bears on the probability of discovery in another area. Rather, we may have a situation where, for example, the value of cost information in area A is zero, if and only if it is known that the cost in area B is always less. Simultaneously, the value of information about area B may depend on information produced about area A. This interaction can be important even if the characteristics of the areas are independent, because the timing of exploration and the efficient sequence of extraction will depend on information produced about both areas.

In the following discussion, I assume that the probability distribution of returns is known, and information serves to improve the process of search. Of course, when population distributions are unknown, we must distinguish information that improves estimates of population distributions from information that improves the process of search when the population distribution is known. However, at least in some restricted cases, Michael Rothschild has shown that the qualitative properties of search are unchanged when the population distribution is unknown.

The problem I will discuss is not offered for its generality. It is a specific and stylized example. However, it does show that when the characteristics of nonreplenishable resources can be screened, it may not be desirable to screen all resources at the same intensity, even though *ex ante* all are identical.

Assume that there are only two types of deposits that can occur at any location, and that they are distributed randomly and in equal proportions. It is convenient, although not necessary, to assume that both types of deposits

* University of California-Berkeley. I am grateful to Partha Dasgupta and Joseph Stiglitz for helpful suggestions.

[1] "Capital Investments of the World Petroleum Industry," The Chase Manhattan Bank, published annually.

have zero extraction cost but differ in unit sampling or discovery cost and are of the same size. Let the cost be C_1 and C_2 per unit of the resource discovered, with $C_2 > C_1$. Furthermore, assume that C_2 is less than the price of a substitute so that both types of deposits will be produced at some time.

Experiments such as aerial surveys, seismic testing and core drillings serve to identify favorable (i.e., C_1) locations. For the purpose of this example, let us suppose that the outcome of any experiment applied to a location is either (+) or (−), and the set of feasible experiments form a continuum indexed by the parameter $\theta \epsilon [0, 1]$. The posterior expected cost of a location conditional on the outcome of an experiment indexed by θ is assumed to be

$$(1) \quad C^{+}(\theta) = \frac{1}{2}(1 + \theta)C_1 + \frac{1}{2}(1 - \theta)C_2$$

if the location is given a (+) or favorable indication, and

$$(2) \quad C^{-}(\theta) = \frac{1}{2}(1 - \theta)C_1 + \frac{1}{2}(1 + \theta)C_2$$

otherwise.

In other words, the experiment symmetrically sorts locations by expected cost with an accuracy measured by the parameter θ. The parameter θ will be called the screening level of the experiment. Note that experimentation serves only a pure sorting function and it is therefore necessary that

$$C^{+}(\theta) + C^{-}(\theta) = C_1 + C_2 = 2C$$

where C is the average cost per unit of resource over all locations and is independent of θ.

Let the cost of applying a screening experiment of accuracy θ to one location be given by $E(\theta)$ Since deposit sizes are assumed identical, $E(\theta)$ can be measured per unit of resource screened.

Define

$$\tilde{C}^{+}(\theta) = C^{+}(\theta) + 2E(\theta).$$

The cost of screening, $E(\theta)$, is absorbed in the cost of producing from the favorably identified locations because these locations must be identified before extraction at cost $C^{+}(\theta)$ can take place. The factor two arises because only one-half of those locations tested will be, on the average, favorably identified.

Let us now examine the allocation of screening effort in a decentralized economy with complete futures markets. Individuals are endowed with an amount S_{0i} of the resource, $i = 1, \ldots, n$, in the form of deposits with the characteristics described previously, and it is assumed that individuals are informed of the nature of the distribution of resource deposits. Suppose that a positive level of screening maximizes discounted net consumer surplus. We will show that this allocation cannot be maintained by a decentralized market economy in which individuals screen their endowment at the same level

Suppose that such an equilibrium existed, defined by $\hat{\theta} > 0$. The prices which sustain the market allocation must vary so that the discounted net rent on those deposits which are produced is constant over the extraction period. Consequently, the trajectory of prices in futures markets would be as shown in Figure 1.

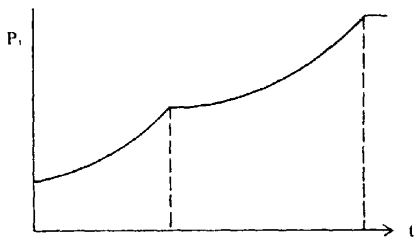


FIGURE 1

At time T_1, the favorably identified locations are exhausted and production shifts to the remaining deposits If there exists a substitute with marginal cost greater than $C^{-}(\hat{\theta})$, it would be introduced at time T_s.

For $t \in (T_0, T_1)$

$$(3) \qquad \frac{dP_t}{dt} = r[P_t - \tilde{C}^{+}(\hat{\theta})].$$

and for $t \in (T_1, T_s)$

$$(4) \qquad \frac{dP_t}{dt} = r[P_t - C^{-}(\hat{\theta})].$$

Let us take the perspective of an individual who must decide when to sell his endowment and whether it is to his advantage to depart from the equilibrium level of screening, $\hat{\theta}$. Let θ_k be the screening level chosen by the individual If $\theta_k = \theta$, then by (3) and (4), he should be indifferent with regard to producing from the better locations at any time between T_0 and T_1, and he should be indifferent with regard to producing from the remaining locations at any time between T_1 and T_s. Therefore, he may as well produce both varieties at T_1. However, if both types are produced simultaneously, there is no advantage in screening. With no screening, average present-value extraction costs are the same, but the cost of screening is avoided.[2] Since this is a competitive market, each individual can reasonably make this calculation and assume his actions will have a negligible impact on prices. However, since the cost characteristics of all endowments are assumed to be identical, if it is in the interest of one person to deviate from the postulated equilibrium screening level, then it must be to everyone's advantage to do so. The collective action will, of course, change prices, and it follows that $\theta > 0$ cannot be a market equilibrium allocation

Since $\hat{\theta} > 0$ induces individuals to choose $\theta_i = 0$, for all i, let us examine $\hat{\theta} = 0$ as a potential market equilibrium allocation With $\theta_i = 0$ for all i, locations are not distinguished and all deposits are produced at the average unit cost C The price trajectory corresponding to $\hat{\theta} = 0$ would yield capital gains on the more costly locations if these were screened. For this price trajectory, screening would be profitable if the cost of sorting locations were sufficiently small If this were the case, all individuals would prefer to screen their endowments Yet we have shown that a uniform positive screening level is not a market equilibrium either. Thus a market equilibrium in which all agents screen at the same level does not exist. Of course, the result is dependent on the assumption of constant marginal costs.

[2]This is possible because there is a market for oil, regardless of its source.

There still remains the possibility of an equilibrium market allocation in which individuals choose different levels of screening. In this case, the allocation can be characterized by a distribution function, $f(\theta)$, where

$$\int_{\theta_1}^{\theta_2} f(\theta)d\theta = F(\theta_2 - \theta_1) \tag{5}$$

is the fraction of the total resource endowment $S_0 = \sum_i S_{0i}$ which is screened in the interval (θ_1, θ_2). The distribution is symmetric in the sense that the amount of the resource with extraction cost $\bar{C}^+(\theta)$ must equal the amount with extraction cost $C^-(\theta)$.

If prices are quoted in futures markets, individuals will choose to produce their endowments at different points in time. At any rate of price increase, the rate of return on deposits is higher the greater is the cost of extraction. Therefore, those with higher screening levels should produce from their better locations earlier, and from the remainder later, than those with lower screening levels.

Let $\hat{H}$ be the set whose elements satisfy $f(\theta) > 0$ It is easy to see that a market equilibrium cannot be sustained if $\inf\{\hat{H}\} > 0$. The argument is the same as that used previously. Let T_1 be the time at which all the positively identified locations are exhausted and extraction shifts to the remaining locations. By doing no screening and producing at all locations at T_1, the value of the endowment is unchanged, but the cost of screening is eliminated. Therefore, $\hat{\theta} = 0$ would be preferred. In other words, for a positive level of screening to be sustained as a market equilibrium, it is necessary that others do less screening. Otherwise, zero screening would be a preferred allocation. For the market allocation to be an equilibrium, it must be impossible to increase the value of an endowment by choosing a different screening level.

In general there does exist a distribution of screening activity that is a market equilibrium.

We will illustrate this for the special case in which there are two alternative screening choices: either screen at level θ_0 or not at all. An equilibrium for this example would require that profits be independent of the choice of screening level, taking prices as given.

If a market equilibrium exists, the price trajectory must be as shown in Figure 2, where

$$(6)\qquad \frac{\dot P_t}{P_t - \bar C^+(\theta_0)} = r \qquad \text{for} \qquad t < T_1$$

$$(7)\qquad \frac{\dot P_t}{P_t - C} = r \qquad \text{for} \qquad T_1 < t < T_2$$

and

$$(8)\qquad \frac{\dot P_t}{P_t - C^-(\theta_0)} = r \qquad \text{for} \qquad t > T_2 .$$

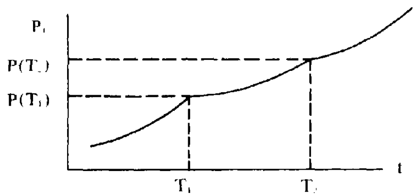


FIGURE 2

It is necessary that

$$(9)\qquad \frac{S_{0i}}{2}[(P(T_1) - \bar C^+(\theta_0))e^{-rT_1} + (P(T_2) - C^-(\theta_0))e^{-rT_2}] = S_{0i}[(P(t) - C)e^{-rt}]$$

for $t \in (T_1, T_2)$, for all $i = 1, \ldots, n$,

and the screening distribution must be such that

$$(10)\qquad E(\theta_0) = \frac{\theta_0}{4}(C_2 - C_1)(1 - e^{-r(T_2 - T_1)}).$$

Time enters explicitly into the determination of the screening equilibrium because it is the separation in time between production of the better and poorer locations that sustains a profitable screening level.

Given the demand correspondence for the resource, the equilibrium screening distribution can be determined. For example, if demand were inelastic and equal to Q_0, then the amount of the resource that is not screened would be

$$(11)\qquad F(0)S_0 = Q_0(T_2 - T_1),$$

and by equation (10), in equilibrium

$$(12)\qquad F(0) = \frac{Q_0}{rS_0}\ln\left[\frac{\frac{(C_2 - C_1)}{2}}{\frac{(C_2 - C_1}{2} - \frac{2E(\theta_0)}{Q_0}}\right]$$

We have assumed throughout this analysis that producers take prices as given. One might argue that the failure of a single-level screening equilibrium stems from the failure of producers to anticipate the consequences of their actions. The analysis shows that this is not the case. An equilibrium clearly does exist but it has the property that those endowments that were identical *ex ante* are differentiated *ex post*.

The stability of such an equilibrium has not been demonstrated. The example showed that Nash behavior could lead to cobweb-type oscillations. It can be shown however, that if the economy is at the differentiated equilibrium, it is stable for small changes in the screening distribution.

The results clearly depend on the specific structure of the economy. For example, the addition of random changes in the market price due to disturbances in the economy could be sufficient to generate a single price equilibrium. The differentiated equilibrium arises in part because agents are well-informed and can predict the consequences of decisions. In particular, future rates of return must be predictable. This could be the case in more complicated scenarios where, for example, the size of the stock is uncertain but costs are predictable.

A final note concerns the general intuitive explanation of the differentiated screening equilibrium. If a concave objective function is maximized, there should be no discontinuities if the solution occurs at an interior point and jumps are avoidable (see Vind). The market equilibrium maximizes a particular function, the sum of consumer and producer surplus. This function is concave in the level of screening if the cost of screening is a convex function. A single level of screening must imply a discontinuity in the objective function when production shifts from the high grade to the low grade deposits. This cannot be efficient. The optimum must entail a distribution of screening intensities, including no screening at some point in time.

REFERENCES

Richard J. Gilbert, "Search Strategies for Nonrenewable Resource Deposits," Institute for Mathematical Studies in the Social Sciences, Technical Report, No. 196, Stanford University 1976.

Michael Rothschild, "Searching for the Lowest Price When the Distribution of Prices is Unknown," *J. Polit. Econ.*, July-August 1974, *82*, 689–711.

Karl Vind, "Control Systems with Jumps in the State Variables," *Econometrica*, *35*, 1967, 273–77.

Nonprice Competition

By MICHAEL SPENCE*

It is difficult to assess the performance of many industries and markets in the private sector without devoting some attention to the nonprice dimensions of competition. I recently had an opportunity to study the refining and marketing segments of the petroleum industry.[1] In recent years, the rates of return in those segments of the industry have not been excessive. In fact, one could make a plausible case that they may have been too low. But the absence of excessive profits, here and elsewhere, is not definitive evidence of adequate performance. It is evidence that prices are not inappropriate given costs. But the costs may have been too low or too high, due to a failure in the area of nonprice competition. In the petroleum industry, analysts have suggested that the historical failure was an excess of nonprice competition in retail outlets, advertising and promotions of a variety of kinds. The retail outlets in particular are thought to have cost the consumer more than they were worth. While the evidence for these propositions is not decisive, a recent dramatic trend downward in the number of retail outlets, stimulated by successful entry and competition from independent retailers, selling with lower prices and higher volumes, is suggestive of prior excess capacity in retailing.

Nonprice competition, for the purposes of this paper, refers to activities by firms or corporations that shift the demands for products, their own and those of their rivals. There are many examples.[2] I mentioned the number and quality of retail outlets. Competition among airlines for passengers has been much discussed in the literature.[3] With a regulated price, airlines compete with numbers of flights and in other ways. Excesses of nonprice competition in this instance are thought to result from a regulated price that may be too high. An apparently successful experiment with no regulation in California and a pervasive belief that competition leads to good results has led to a movement for deregulation nationally.

Policy prescription and the assessment of market performance is hampered in these and other areas by the absence of models that deal explicitly with the interaction of demand and cost, price and nonprice competition, and entry.

I. A Model

To capture the interaction of price and nonprice competition, it is necessary to specify the demands for a collection of products which are imperfect substitutes for each other. The demands depend upon prices or quantities and upon nonprice activities. Let the quantity of good i be x_i and the level of a nonprice activity by firm i be a_i. The demands are equivalent to and derivable from the gross benefits generated by the products. The gross benefits depend upon $X = (x_1, \ldots, x_n)$, and $A = (a_1, \ldots, a_n)$. For this paper I have assumed that the benefits have the form

$$(1) \qquad B(X, A) = G\left[\sum_{i+1}^{n} \phi_i(x_i, a_i)\right],$$

where $G(s)$ is a concave function of s, and $\phi_i(x_i, a_i)$ is concave in x_i and has the property $\phi_i(0, a_i) = 0$. In addition $G(0) = 0$.

The implied inverse demands are found by taking the derivatives of B with respect to x_i. The reason is that consumers act so as to maxi-

*Harvard University. This research was supported by the National Science Foundation, grant SOC 76-16827. I am indebted to Paul Joskow and the staff of the Royal Commission on Petroleum Products Pricing in Ontario, for helpful comments.

[1]The Royal Commission of Petroleum Products Pricing, discusses the state of nonprice competition in this industry.

[2]One example is new product competition, or monopolistic competition. It is discussed in E. L. Bishop, Avinash Dixit and Joseph Stiglitz, and Spence (1974). The present analysis is related to the problem of a monopolist setting product quality, discussed in Eytan Sheshinski, and Spence (1975).

[3]A discussion of this problem is given by J. Panzar

mize $B(X, A) - pX$, where $p = (p_1, \ldots p_n)$ are prices.

The inverse demands are

$$(2) \quad p_i(X, A) = G'(s) \frac{\partial \phi_i}{\partial x_i}(x_i, a_i),$$

where

$$s = \sum_j \phi_j(x_j, a_j).$$

In this paper, I want to confine myself to symmetric cases. Let (x, a) be the quantity, and nonprice expenditure of the representative firm, and let $\phi(x, a) = \phi_i(x, a)$ for all i. The costs of the representative firm are $c(x, a)$. The number of products or firms is n. It follows that $s = n\phi(x, a)$.

The total surplus is

$$(3) \qquad T = G(s) - nc(x, a)$$

and it can be written in the form

$$(4) \qquad T = G(s) - s\,\frac{c(x, a)}{\phi(x, a)},$$

using the fact that $n = s/\phi$. The surplus is maximized when c/ϕ is minimized with respect to x and a, and when the numbers of products are expanded so that

$$(5) \qquad G'(s) = \min_{(x, a)} \left(\frac{c}{\phi}\right).$$

This result can be rationalized by noting that the contribution to the surplus of the last product is $G'\phi$. Thus $\frac{c}{G'\phi}$ are the costs per dollar of benefits generated by the last product. These are minimized with respect to x and a. And the numbers are increased until marginal costs per dollar of benefits are equal to one.

II. Profits and the Equilibrium

The profits of the representative firm are

$$(6) \qquad \pi = G' x\phi_x - c.$$

Firms maximize these with respect to x and a, and then entry occurs until profits are driven to zero, or returns to normal. Entry is crucial in this process. It forces firms to adopt strategies that minimize $\frac{c}{x\phi_x}$, a term is proportional to the costs per dollar of revenue for the representative firm. This is what one would expect from the market. Firms respond to a threat to survival posed by entry by reducing the costs per dollar of revenue as much as possible. The equilibrium occurs when

$$(7) \qquad G'(s) = \min_{(x, a)} \left(\frac{c}{x\phi_x}\right)$$

These relatively simple characterizations of the equilibrium and the optimum permit a comparison of the two.

Note first that the equilibrium is as if the market were maximizing $G(s) - s\left(\frac{c}{x\phi_x}\right)$. Because ϕ is concave in x (i.e., demands are downward sloping) $x\phi_x < \phi$. Two things follow immediately. The minimum of $\frac{c}{\phi}$ is less than the minimum of $\frac{c}{x\phi_x}$. Thus profits are negative at the optimum. And gross benefits, $G(s)$, are smaller at an equilibrium than at the optimum.

The remaining differences have to do with differences induced by the minimization of $\frac{c}{\phi}$ and $\frac{c}{x\phi_x}$. It is useful to examine these first for a particular case. Let us assume that $\phi(x, a) = A(a)x^{\alpha(a)}$. For this case, the inverse demand is

$$p = G' A \alpha x^{\alpha - 1},$$

and thus $1 - \alpha$ is the quantity elasticity of the inverse demand. Under these circumstances

$$x\phi_x = \alpha\phi$$

It then follows that

$$\left(\frac{c}{x\phi_x}\right) = \left(\frac{c}{\alpha\phi}\right)$$

Therefore the optimal and equilibrium quantities given a, are the same. Let

$$m(a) = \min_x \left(\frac{c}{\phi}\right) \tag{8}$$

Then for this case

$$r(a) = \min_x \frac{c}{x\phi_x} = \frac{m(a)}{\alpha(a)} \tag{9}$$

At an optimum, $m(a)$ is minimized with respect to a, so that $m'(a) = 0$. At an equilibrium m/α is minimized, so that

$$\frac{m'}{m} = \frac{\alpha'}{\alpha} \tag{10}$$

Suppose $\alpha'(a) > 0$, so that the nonprice activity reduces the quantity elasticity of the inverse demand. Roughly, the nonprice competition makes the inverse demands flatter, however much or little it shifts them. If $\alpha' > 0$, then $m'(a) > 0$ at an equilibrium, and the equilibrium level of a is above the optimum. The converse also holds. If $\alpha' < 0$, then $m'(a) < 0$ at an equilibrium and the level of a is too low.

Figure 1 depicts the equilibrium and the optimum for the case $\alpha' > 0$.

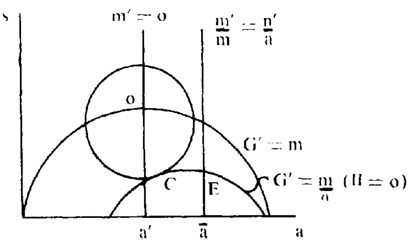


FIGURE 1

The equilibrium is at E, the optimum at 0. In the special case where $\alpha'(a) \equiv 0$, the equilibrium and the optimum have the same level of a: it satisfies $m'(a) = 0$.

III. The Constrained Optimum

Because profits are negative at the optimum, it is of interest to know how well one can do with profits constrained to be nonnegative. For the case, $\phi = Ax^{\alpha}$, the answer can be seen easily in Figure 1. The line $G' = \frac{m}{\alpha}$ is the zero profit line. Isototal surplus contours are vertical through $G' = m$, and horizontal through $m'(a) = 0$. Thus for $\alpha' > 0$, the constrained optimum occurs at a point like C. The level of a is below the equilibrium level. So also is the level of s, and therefore the gross benefits. This means that the industry incurs excessive costs, as a result of excessive nonprice competition. There may be too many or too few products. For reasons of space, I shall not draw the figure for the case $\alpha' < 0$. In that case a is lower at the equilibrium than at either optimum. Relative to the constrained optimum, gross surplus is too high at an equilibrium. The competitive industry incurs excessive costs, not from nonprice expenditures directly, but from excessive entry. In that case, the number of products is too large. The following proposition summarizes these results.

PROPOSITION 2: If $\phi(x, a) = A(a)^{\alpha(a)}$, then

(i) quantities at the equilibrium and the optimum are the same.

(ii) If $\alpha'(a) > 0$, the equilibrium has more nonprice competition than the optimum and the constrained optimum; and while the gross benefits are higher at the equilibrium than the constrained optimum, costs are too high, partly due to excessive nonprice competition.

(iii) If $\alpha'(a) < 0$, there is a deficiency of nonprice activity relative to both the optimum and the constrained optimum; gross benefits are higher at the equilibrium than at the constrained optimum, but costs wipe

out these benefits in the form of excessive entry of new products.

(iv) If $\alpha'(a) = 0$, the levels of a are the same at the equilibrium, the optimum and the constrained optimum. The equilibrium and the constrained optimum are identical.

It is interesting to note that none of these results depends upon the properties of the cost function. They are therefore consistent with a wide range of possible ways in which the nonprice activity affects costs. It can increase or decrease marginal costs, or leave them unchanged. Or, as in the airline case, the effect on marginal costs of increased flights can be lumpy and discontinuous.

IV. The General Case

The relation between the equilibrium and either optimum is determined by two relationships. The first order conditions for a minimum of $\frac{c}{\phi}$ are

$$\frac{c_a}{c} = \frac{\phi_a}{\phi} \tag{11}$$

and

$$\frac{c_x}{c} = \frac{\phi_x}{\phi}$$

while for an equilibrium, the conditions for a minimum of $\frac{c}{x\phi_x}$ are

$$\frac{c_a}{c} = \frac{\phi_{xa}}{\phi_x}$$

and

$$\frac{c_x}{c} = \frac{1}{x} + \frac{\phi_{xx}}{\phi x}.$$

For a given x, the optimal a is above the equilibrium when

$$\frac{\phi_a}{\phi} > \frac{\phi_{xa}}{\phi_x}$$

and conversely.

This says that the elasticity of $\frac{G'\phi}{x}$, the average value of the marginal product (averaged that is, over purchases) with respect to a, exceeds the elasticity of $G'\phi_x$, the value of the product to the marginal consumer.

This is the effect which results from the difference between the average and marginal purchaser. It is the elasticity effect described earlier.

For a given a, the optimal x exceeds the equilibrium x if (from the conditions above)

$$\frac{x\phi_x}{\phi} > 1 + \frac{x\phi_{xx}}{\phi_x}$$

and conversely.

Because

$$\frac{\partial}{\partial x}\log\left(\frac{x\phi_x}{\phi}\right) = \frac{1}{x} + \frac{\phi_{xx}}{\phi} - \frac{\phi_x}{\phi}$$

This condition can be written

$$\frac{\partial}{\partial x}\log\left(\frac{x\phi_x}{\phi}\right) < 0$$

Therefore, for a given a, the equilibrium quantity falls short of the optimum if revenues $G'x\phi_x$ are declining fractions of the incremental benefits $G'\phi$, and conversely. In the special case $\phi = Ax^{\alpha}$,

$$\frac{x\phi_x}{\phi} \equiv 1 - x\frac{\phi_{xx}}{\phi_x} = \alpha.$$

Thus the result that quantities are optimal given the level of a, at the equilibrium. These two effects combine to determine the relation between the equilibrium and the optimum. Figure 2 shows one case, where

$$\frac{\partial}{\partial a}\left(\frac{x\phi_x}{\phi}\right) > 0$$

and

$$\frac{\partial}{\partial x}\left(\frac{x\phi_x}{\phi}\right) > 0.$$

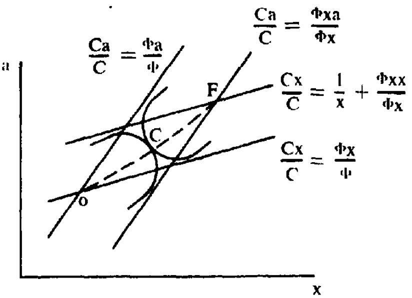


FIGURE 2

Here both x and a are too high at an equilibrium. Of the four possible cases, two are formally ambiguous.

The constrained optimum involves maximizing $T = G - \frac{sc}{\phi}$ subject to $G' \geqq \frac{c}{x\phi_x}$. Let λ be the shadow price of the constraint. The Lagrangian for the problem is

$$L = G + \lambda s G' - s\left[\frac{c}{\phi} + \frac{\lambda c}{x\phi_x}\right].$$

Therefore the constrained optimum includes the (x, a) that minimizes

$$\frac{c}{\phi} + \lambda \frac{c}{x\phi_x}$$

a linear combination of the quantities that are minimized in the equilibrium and the optimum. The constrained optimum therefore lies on the locus of the tangencies of the contours $\frac{c}{\phi}$ = a constant and $\frac{c}{x\phi_x}$ = a constant. This is the dashed line in Figure 2. The constrained optimum is a point like C in Figure 2. Its relation to the equilibrium is similar to that of the optimum.

All of the above can be summarized by noting that market performance is determined by whether the ratio of revenues to incremental surplus increases or declines in x and a. This ratio measures the extent to which social benefits are captured in the form of revenues. A market tends to oversupply services that increase this ratio, and to undersupply those which do not. This is particularly easy to observe in the constant elasticity case, because the ratio does not depend on the quantity sold. Economists are accustomed to thinking that firms with market power prefer low price elasticities of demand. The preceding results appear to contradict this. But the question is what is hypothetically being held constant when the elasticity is reduced. Firms prefer large demands too, and one must be careful not to confuse that with low elasticities. If one holds the gross benefits constant, then firms that cannot price discriminate prefer high elasticities because the ratio of revenues to gross benefits rises with the elasticity. That is all that the preceding analysis requires. There may be other respects in which firms "prefer" low elasticities, but they will not contradict this proposition, nor the implied consequences for market performance.

REFERENCES

E. L. Bishop, "Monopolistic Competition and Welfare Economics," in R. Kuenne (ed.) *Monopolistic Competition Theory*, New York 1967.

Avinash Dixit, Joseph E. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," Stanford, 1974.

J. Panzar, "A Model of Regulated Oligopoly with Product Quality Variation: The Case of Passenger Airlines," Center for Research in Economic Growth, Stanford University, 1973.

Eytan Sheshinski, "Price, Quality and Quantity Regulation in Monopoly Situations," *Economica*, May 1976.

Michael Spence, "Product Selection, Fixed Costs, and Monopolistic Competition," *Rev. Econ. Stud.*, forthcoming, 1974.

———, "Monopoly, Quality, and Regulation," *Bell J. Econ.*, Fall 1975.

The Ontario Royal Commission of Petroleum Products Pricing, June 1976.

APPLICATIONS OF MICROSIMULATION METHODOLOGY

Does Your Probability of Death Depend on Your Environment? A Microanalytic Study

By GUY H. ORCUTT, STEPHEN D. FRANKLIN, ROBERT MENDELSOHN AND JAMES D. SMITH*

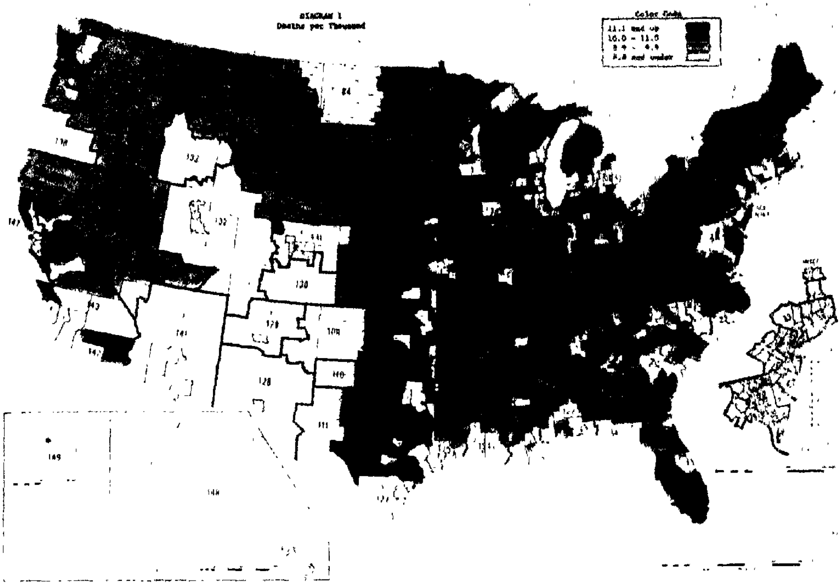
There is a growing interest in detecting manageable environmental changes which would improve overall health and life expectancy. The work we report explores differences in mortality rates in county groups (as defined by the Census) to discover whether environmental differences among them affect death probabilities. Casual inspection of Diagram 1 offers evidence that the probability of dying is significantly different for individuals in different parts of the country. However, these variations in mortality rates are not necessarily the result of environmental factors which differ across the country; they could also be the result of different personal characteristics of local populations For example, an area with an unusually wholesome environment may have a high death rate because it possesses relatively more elderly citizens. The primary objective of this paper is to remove the influence of personal characteristics so that the effects of area specific factors can be detected

*Professor of economics, Yale University, research associate, The Urban Institute, Ph D candidate, Yale University; visiting professor, Yale University, respectively The research for this paper was conducted with financial support from the National Science Foundation to The Urban Institute, grant No SOC73-05420-A01 for the "Simulation of the Distribution of Income." Substantial support in the form of staff time, computing, housing and secretarial services was supplied by the Institution for Social and Policy Studies of Yale University Valuable assistance in preparing this paper was received from Amihai Glazer and Jan Stolwijk. The views expressed are those of the authors and do not necessarily represent the views of the National Science Foundation, The Urban Institute or Yale University

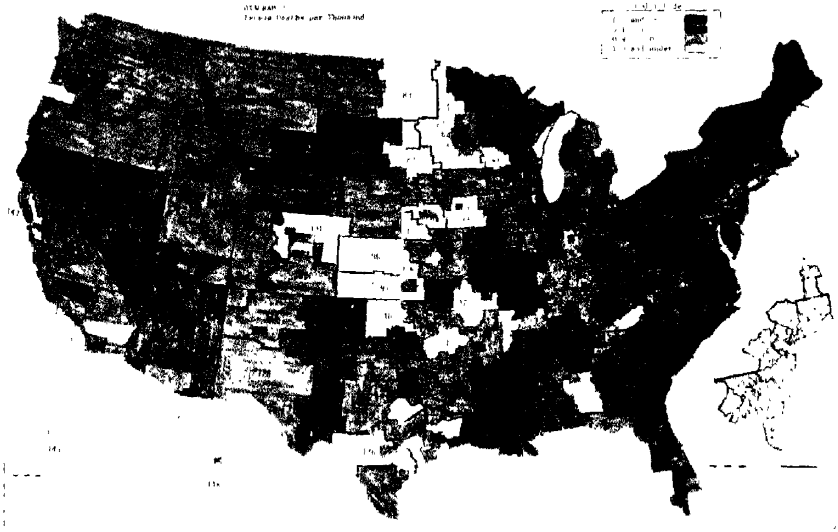
I. Research Strategy and Data Base

Unfortunately, the absence of a micro data base providing extensive information about decedent's lifetime characteristics complicates the analysis of mortality rates We were able to take advantage of three large data sets which, when combined, provide somewhat comprehensive information about individuals and their environment. Information from the two million recorded death certificates filed in 1970 were organized by county group into 28 race, sex, and age cells. The second body of data, also about two million observations, was the 1970 Census of Public Use Sample, which provides geographic identification down to the county group The Public Use Sample provided information about the living population in race, sex, and age cells within each of the 405 county groups. Macro variables describing the entire population of the county groups were also computed from the Public Use Sample. Finally, the machine readable 1970 City and County Data Book provided additional information about the overall characteristics of county groups.

The basic research strategy was to attribute as much as possible of the between-county group variation of death rates to person specific factors. Since spatial features of residual variation of death rates appears nonrandom between county groups, the existence of significant environmental influences on death probabilities are strongly suggested The residual variation might reasonably be explained by area specific factors such as climate, industrialization, and economic vitality. Though we have not yet been able to



Deaths per Thousand
Color Code
11.1 and up
10.0 - 11.0



and under

include many appropriate measures of such factors, we did supply several variables which hopefully serve as proxies for the omitted information. On this basis an attempt was made to relate variation between county groups within race, sex, and age brackets to area specific factors jointly with person specific factors.

II. Removing Effects of Person Specific Factors

To take advantage of the detailed micro information in the Public Use Sample, a microanalytic simulation model, *DYNASIM*, was used to simulate death.[1] The death module of *DYNASIM* generated the expected number of deaths for each county group based on national mortality probabilities for 1970 by age, race, sex, marital status, education, and children ever born (for women). The difference between the actual death rates (from the death certificate tapes) and the simulated death rates is shown for each county group in Diagram 2. Note the differences between Diagram 2 and Diagram 1. The high recorded death rates in the center of the country disappear when the model removes the influence of the personal characteristics. On the other hand, though the actual death rates in the Southeast appear to be low, the model indicates they are relatively high once personal factors are considered.

In order to test new personal variables and area specific terms, we developed a regression model of the following form:

$$(1)\qquad D_i = B_o + B_1X_{1i} + B_2X_{2i} + \ldots + \gamma_1Z_1 + \gamma_2Z_2 + \ldots + E_i$$

Whether the individual died over the years (D_i: a zero-one dummy) is equated to his set of personal characteristics ($X_{1i}, X_{2i} \ldots$), the variables for his county group ($Z_1, Z_2, \ldots$) and an error term (E_i). In order to estimate this relationship, the individual observations are aggregated by cell (age, race, and sex) to the county group level.

$$(2)\qquad \sum_1^N D_i = NB_o + B_1\sum_1^N X_{1i} + B_2\sum_1^N X_{2i} + \ldots + N\gamma_1Z_1 + N\gamma_2Z_2 + \ldots + \Sigma E_i$$

Dividing equation (2) by N, the number of observations in each cell and each county group, yields the regression model we estimated for the probability of death (P_D).

$$(3)\qquad P_D = \frac{\sum_1^N D_i}{N} = B_o + B_1\frac{\sum_1^N X_{1i}}{N} + B_2\frac{\sum_1^N X_{2i}}{N} + \ldots + \gamma_1Z_1 + \gamma_2Z_2 + \ldots \frac{\sum_1^N E_i}{N}$$

Equation (3) was subsequently weighted by the square root of the populations in each cell in order to account for heteroscedasticity. We then fit equation (3) to each age, race, and sex cell. One of the advantages of examining each of these relatively homogeneous cells separately is that it should lessen the importance of potential interaction terms. The linear form of Equation 3 may be a justifiable approximation given the stratification of the data.

The first set of regressions incorporated the personal characteristics already included in *DYNASIM* and added relative income as a new variable.[2] Two of these regressions are shown in columns 2 and 9 of Table 1 for whites aged 45–64. Despite the narrow classification of these individuals by race, age, and sex, the signifi-

[1]For a full presentation of the concept and application of microanalytic modeling and simulation as well as of the research behind *DYNASIM*, see *Policy Exploration Through Microanalytic Simulation*, by Orcutt, Steven B. Caldwell, Richard Wertheimer II, and Franklin, Gary Hendricks, Gerald E. Peabody, Smith, Sheila Zedlewski. Urban Institute, September 1976, Washington, D.C.

[2]Relative income is family money income divided by the official poverty score attaching to a family's characteristics.

TABLE 1—MULTIPLE REGRESSIONS FOR EACH CELL ON DEATH RATES PER THOUSAND PERSONS USING COUNTY GROUP AND PERSONAL VARIABLES

	White Males				Black Males			White Females				Black Females		
	25–44	45–64	45–64	65+	25–44	45–64	65+	25–44	45–64	45–64	65+	25–44	45–64	65+
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Personal Variable														
Constant	3 86	−17 13	13 2	78 9	9 4	64.2	241 6	− 45	− 281	11 6	61 5	4 56	32 7	194.8
Age	03	74	36	54	− 19	20	− 65	04	12	− 08	32	− 09	.26	−.57
	(55)	(3 57)	(1 64)	(91)	(1.70)	(66)	(1 22)	(1.47)	(1.26)	(81)	(75)	(1 47)	(1.34)	(1.22)
Schooling	− 21	− 72	09	− 34	−.48	− 32	06	− 08	− 04	−.08	−1 26	− 19	− 02	−.51
	(2 99)	(4 97)	(29)	(46)	(3.06)	(83)	(06)	(1 41)	(59)	(47)	(2 35)	(1 86)	(.10)	(.79)
Relative Income	15	− 44	− 11	−2 34	17	−1 16	− 96	− 01	− 12	− 17	−3 19	− 13	−1 69	−3.75
	(97)	(1 79)	(19)	(1 42)	(56)	(1 26)	(41)	(13)	(.97)	(1 47)	(3 03)	(65)	(3 36)	(2.00)
% of Cell Widowed	1 41	43 94	15 8	−1 41	−1 14	−6 81	1 16	−.99	8 99	5 02	20 2	.79	−4 46	3.11
	(22)	(4 30)	(1 76)	(.12)	(14)	(84)	(.11)	(.65)	(4 47)	(2 08)	(2 68)	(37)	(1 53)	(.42)
% of Cell Divorced	− 82	37 18	12 5	34 09	−3 03	−10 98	−17.0	.04	9 83	− 10	40 2	3 15	−3.39	−1.89
	(.40)	(7 57)	(1.90)	(1 66)	(1 29)	(1 76)	(1 19)	(04)	(5.19)	(03)	(2 30)	(2 91)	(1.17)	(.16)
% of Cell Never Married	3 84	7 32	−7 6	11 38	−3.94	−4 00	−24 3	62	8 74	2.81	23 2	1.27	88	9.95
	(2 97)	(1 86)	(1 32)	(84)	(2 42)	(66)	(2 10)	(79)	(4.74)	(97)	(1 97)	(1 01)	(.17)	(.83)
Avg No Kids per Female in Cell								01	− 17	49	−2 90	23	.43	1.20
								(07)	(92)	(1.29)	(2 68)	(1 33)	(1.09)	(1.05)
County Group (CG)														
CG Age in Years	01		− 27	−2 06	26	− 27	−1 22	01		− 02	− 83	14	−.24	−.75
	(.46)		(3 46)	(7 02)	(2 25)	(74)	(1 12)	(1.55)		(.44)	(3 74)	(2 24)	(1.24)	(98)
CG Schooling in Years	10		− 78	−2 63	− 41	−1 29	−4.39	06		− 26	−1 27	01	−.88	−1.84
	(1 17)		(1 93)	(2 05)	(−1 21)	(1 25)	(1 43)	(1.20)		(1 13)	(1 40)	(04)	(1.59)	(.84)
CG Relative Income	− 10		−1 74	.65	80	−6 98	1 08	.01		−1.03	4.35	94	−4.37	−3.20
	(58)		(2 85)	(23)	(65)	(1 77)	(09)	(12)		(1 91)	(1 89)	(1 38)	(2.04)	(.38)
CG % 18 & Over Widowed	−.49		−1 32	4 97	22	−3 04	−9 82	− 12		06	2 54	−.45	−1.84	−6.12
	(2 81)		(2 17)	(2.24)	(32)	(1 44)	(1 67)	(1 25)		(15)	(1 81)	(1.24)	(1.69)	(1.48)
CG % 18 & Over Divorced	1.78		47 3	269 6	6 30	68 4	37 13	− 22		11.75	57.08	−14 6	35.4	−91.7
	(75)		(5 21)	(6 91)	(44)	(1 55)	(27)	(17)		(2.28)	(2.00)	(1 82)	(1.43)	(.97)
CG % 18 & Over Never Married	7 82		24 1	23 7	26 5	−3 18	98.7	7.14		8 66	−29 06	5 60	−3.74	6.66
	(3 25)		(2 80)	(74)	(2.44)	(09)	(.93)	(5 62)		(1 97)	(1.30)	(95)	(20)	(.09)
CG Avg No of Kids (Per Female 18 & Over)	−5 96		14 6	46 6	1.31	−4 10	38 8	81		7 33	29 58	2.21	11.35	20.8
	(2 89)		(2 13)	(1 83)	(15)	(14)	(46)	(69)		(2 04)	(1.64)	(48)	(77)	(.35)
Density (persons per Sq Mile)	00		000	002	001	001	− 001	− 000		000	.001	000	.000	−.001
	(1 42)		(90)	(2 27)	(1 84)	(1 27)	(21)	(.34)		(1.38)	(1.85)	(.60)	(.76)	(.31)
Unemployment Rate	01		04	48	−.28	− 79	−1 91	− 02		17	21	− 21	−.58	−.88
	(95)		(69)	(1 65)	(2 26)	(2 04)	(1 61)	(1 85)		(5 39)	(1.08)	(3 10)	(2 74)	(1 03)
% Owner Occupied Units	− 00		01	18	.03	.11	.11	002		01	−.01	00	.10	−.07
	(72)		(1 13)	(2 76)	(1 35)	(1 32)	(44)	(80)		(1 80)	(12)	(01)	(2 30)	(.42)
% of Units Without Plumbing	02		01	− 05	− 01	− 19	− 42	−.001		− 03	06	− 01	− 20	.05
	(2 64)		(47)	(53)	(.26)	(1 72)	(1 27)	(35)		(2 78)	(48)	(23)	(3 43)	(20)
% Overcrowded Units	02		− 10	−1 05	.11	−.06	− 96	01		.01	− 88	02	04	− 97
	(2 09)		(2 19)	(5.23)	(1 40)	(25)	(1 27)	(1 26)		(20)	(6.21)	(36)	(.28)	(1.89)
R^2	60	41	64	34	22	.14	.10	37	26	40	40	21	.21	.09
SEE	38	1 76	1 40	6 54	2.66	8 41	25 13	20	80	.728	4 25	1 47	4.54	17.25

cance of the coefficients in both regressions suggests that additional personal variables have a role in explaining between county variations of mortality rates. Older people, even within these age brackets, have a greater chance of dying and both more schooling and higher relative income are correlated with lower death rates. A somewhat surprising result is the size and significance of the coefficients of marital status.

III. Evidence of Environmental Effects

Despite our effort to explain between county group variations in mortality rates by means of several personal variables, county group-wide variables are still important when included.

The additional regressions in Table 1 suggests that environmental or area specific variables have a significant impact on mortality rates after controlling for personal variables. Several area specific variables are significant. For instance, independently of their own education, individuals generally live longer in county groups which have more educated citizens. On the other hand, inhabitants of areas which contain many single people (never married) generally tend to have higher death rates. These variables relating to the county group are presumably proxies for other social or economic factors which reduce the probabilities of death. Of course, not all environmental variables have the same effect on every cell. For example, living in areas where the average age of the population is high appears to be deleterious to younger adults and beneficial to the more elderly. Similarly, high unemployment rates are correlated with high death rates for whites but low death rates for blacks.

Despite the evidence in Table 1 that environmental factors are important, identifying the effects of specific environmental factors remains a challenge. Not only are some area specific variables serving as proxies for omitted environmental variables, but also some of the person specific variables may be proxies for environmental effects It is worth noting that the personal variables which were significant in Table 1, are much less significant when county group variables are included. Though we have been able to show that environmental variables are probably important, identifying which area-wide variables should be included and measuring exactly how important they are requires additional research.

Macroeconomic Effects of a Humphrey-Hawkins Type Program

By Barbara R. Bergmann and Robert L. Bennett*

The provisions of the Humphrey-Hawkins bill are still in a state of evolution at this writing, but the body of ideas and value judgements behind the legislative effort going on under that title can probably be summed up by three principles:

1) It should not be deliberate policy of the United States to induce or tolerate high unemployment rates for the express purpose of curing or preventing inflation. (High unemployment rates are defined as rates above some maximum, the most frequently discussed maximum is in the 3.5–4.5 percent range)

2) The government should have an ongoing program to prevent unemployment from exceeding the maximum insofar as possible, and to expeditiously reduce the unemployment rate when it exceeds the maximum for any reason

3) A part of the antiunemployment program should be the provision of specially created jobs on the public payroll for persons who cannot be absorbed into regular jobs in the private or public sectors

Proponents of the Humphrey-Hawkins approach tend to emphasize the economic waste and the human deprivation attendant on high rates of unemployment, and would fight inflation by means other than high unemployment rates Critics of the Humphrey-Hawkins approach include those who are reluctant to give up the option of fighting inflation through deliberately created unemployment or through the toleration or slow cure of whatever unemployment shows up fortuitously.[1] The critics also include those who fear that a public jobs scheme would be difficult to administer, and would prove to be a permanent, unproductive and expensive addition to the public payroll; thus they would prefer to fight unemployment by other means.[2]

In this paper, we formulate a concrete program which we consider to be in the spirit of Humphrey-Hawkins, and assess its budgetary costs and some of its benefits. As an instrument in the assessment of the program, we use an entirely micro-simulated macroeconomic model of the *U.S* economy currently under development by the authors. With the help of this model, we rerun the history of the 1973–75 period under the assumption that either of two versions of a Humphrey-Hawkins program was in effect. Our computer simulation, elaborate as it is, is not set up to detect problems in management or in the motivation of participants. We must make assumptions about how people will act in the context of such a program, and then on the basis of those assumptions assess what differences such a program would have made to unemployment rates, to the fiscal position of federal and nonfederal governments, to the rate

*Department of Economics, University of Maryland Conversations with Charles Brown, Barry Chiswick, Alfred Tella, Alan Fechter and Christopher Clague were helpful, although they bear no responsibility. The Computer Science Center of the University of Maryland supported this research by contributing computer time, and funds from an OEO grant supported part of the personnel costs

[1]A statement of this attitude, although not an outright endorsement of it appears in a letter by Leonard A Lecht of the National Industrial Conference Board, reproduced in a report of the National Commission for Manpower Policy (pp 190–191) For a somewhat equivocal view of this matter, see the testimony of Charles L Schultze and compare pp 2 and 14

[2]See letter of Rudolph Hale of the National Association of Manufactures in the National Commission volume, pp. 198–200 In the same volume, Sar A. Levitan, hardly a foe of job creation, sees definite limits in the number of jobs which could be made available, presumably on the grounds that useful work could not be set up for a larger number If a program of the type we simulate here were set up, this issue would have to be faced seriously

of price change, to *GNP*. Our intention is thus to make a limited but constructive contribution to the debate as to the general desirability of the Humphrey-Hawkins approach, and to the deliberations concerning the concrete form which programs under this approach might assume.

I. An Implementing Program for Humphrey-Hawkins

The program format implementing the Humphrey-Hawkins concept which we have chosen for our simulation experiments has the following characteristics: 1) Standards of eligibility are set up for admission to public service employment *(PSE)* and for continuance in the program in terms of the person's unemployment history. 2) A wage schedule for *PSE* is set up, in terms of some fraction of what the person's wage would currently be in regular employment based on occupational history and an assessment of skill level, up to some maximum. 3) Everyone who wants a public service job and who meets the eligibility requirements is given one, at the prescribed wage, regardless of the state of the economy.

For example, one variant of such a program format which we have simulated runs as follows: Anyone who has been unemployed ten weeks or more and who wants a public service job would be given one at a wage of 75 percent of his presumed wage in a regular job, up to a maximum of $150 per week, with no limit of time on his continuance in the program. (We must acknowledge here that it is considerably simpler in a computer simulation than it is in reality to determine what a person's wage in a regular job would be, and to pay different *PSE* wages to different individuals.[3])

The format of the Humphrey-Hawkins implementation program we have chosen to study does not guarantee in advance any particular level of unemployment and in this sense violates at least some versions of the bill. However, it does embody the idea of an automatic and a relatively quick response to worsening employment conditions regardless of the rate of inflation; it addresses the plight of those most acutely affected by unemployment by directly providing them with employment. The wage feature of our implementation program is counter to the desire of some of the bill's proponents that "prevailing wages" be paid to those on the program. However, it seemed obvious to us that if public service jobs were provided at a wage equal to that prevailing in regular jobs the fears of the opponents of the Humphrey-Hawkins approach would be realized in terms of the size, expense and growth of the program. Under a "prevailing wage" regime private employers would be unable to attract people who had been provided with work by the program. For these reasons, we viewed the "prevailing wage" variant of the program as unlikely of adoption, and have not included it in our evaluative calculations.[4] Given our assumptions concerning eligibility and wages, we have assumed that no one would consider leaving a regular job in the private sector for the express purpose of shifting to a *PSE* job, despite the nonpecuniary attractions some forms of *PSE* might have for some members of the work force.

It is not the intention of proponents of the Humphrey-Hawkins approach that all or even a major part of the burden of fighting unemployment be placed on the provision of public service jobs; the usual monetary and fiscal instruments are to continue to be employed, perhaps more vigorously, skillfully and presciently than at present if that be possible. Obviously, the more that is done by the conventional instruments, the less needs to be done by *PSE*. In our simulation of the 1973–75 period, we have assumed that tax rates and government purchases un-

[3] See the Schultze testimony on the actual dispersion of wages in low wage jobs. The program we have simulated might be viewed as a compromise between Arthur Burns' suggestion that government jobs be provided at below the minimum wage, and the "prevailing rate" wage specification of the bill. See Michael Barth on this and other issues.

[4] This is not to say that a program which took certain individuals onto the public payroll at wages higher than they could get from private employers for the purpose of changing the distribution of income by race, sex or I.Q. would be without merit. See Lester Thurow. Richard Nathan in the National Commission volume (p. 114) considers such an approach "infeasible."

connected with the proposed *PSE* program[5] remained what they in fact were, and that our proposed *PSE* program was added on top of them. It is assumed that the monetary authority accomodated the additional issuance of government debt by acting on reserves to keep interest rates within the announced target range. One of the interesting issues raised by the Humphrey-Hawkins approach for those who accept its basic premises is the amount of reliance which should be placed on *PSE* as opposed to the other instruments. As will be seen, the answer which many currently favor, namely, "as little as possible," is not necessarily the most conservative answer.

In order to simulate the operations of such a program, certain assumptions about the nature of the setup and behavior of the persons and institutions involved must be made explicit. We have chosen to make the following assumptions:

1) Unemployed persons eligible for unemployment insurance *(UI)* will tend to want to go on *PSE* less often than those not eligible. We have set the probability that a person will want to shift to *PSE* from *UI* in a given week at 1 for those with fewer than 10 weeks of *UI* entitlement remaining, and at 1.0 for those with no weeks remaining.

2) Persons on *UI* and on *PSE* will continue to search for regular jobs, but will accept 75 percent as many job offers as persons eligible for neither.

3) State and local governments will reduce the number of their regular employees by two persons for every five persons on *PSE*, up to a maximum of 10 percent of their regular employees.[6]

4) In addition to paying the salaries of the public service workers, the federal government will purchase fourteen cents worth of work materials, training services, and the like from the private sector for each dollar of *PSE* wages, and will hire five additional regular federal workers for each 100 *PSE* workers to administer the program.

5) Welfare payments will decrease by ten cents for every dollar of *PSE* wages paid.

II. A Microsimulated Model of the Macroeconomy

The vehicle we have used to simulate the effects of a Humphrey-Hawkins program is a specially prepared version of the Transactions Model—a microsimulated model of the *U.S.* economy.[7] The Transactions Model represents the *U.S.* economy by a much smaller-scaled simulated economy, in which the actors are 800 worker-consumer-asset holders, 12 firms, each of which produces the product of a particular industrial sector, the federal government, a consolidated state/local government, and the monetary authority. The situation and history of each of the actors in the model is kept track of in considerable detail, and the "action" in the economy consists of decisions by the actors based on their individual situations and on interactions among the actors based on those decisions. All macroeconomic magnitudes generated endogenously by the model are built up on the basis of actions by individuals; in particular, the simulated *GNP* accounts and flow of funds accounts are based on the summed transactions between individual actors, in which money is exchanged against goods, services or claims. The supply side of the economy is fully represented; firms make production and pricing decisions endogenously, and try to hire enough workers to realize their production plans at wages they set endogenously.

The Transactions Model is a particularly apt vehicle for the simulation of the operations of a Humphrey-Hawkins type of program because

[5]There were in fact some *PSE* programs within the period, but we have ignored their existence for present purposes and assumed in effect that the enrolees were in regular federal jobs.

[6]This is on the high side of the range thought reasonable by Fechter (p. 17) and the Congressional Budget Office. Also, see Michael Wiseman on this and other issues.

[7]For a description of an earlier version of the Transactions Model see Bergmann. A more complete version will appear in Bennett and Bergmann

account can be taken of each unemployed worker's situation: the length of time he has been unemployed, his eligibility for *UI*, his customary occupation, the current value of his assets. A further feature of the model which makes it particularly useful in this context is its time frame: although the model produces monthly simulated unemployment rates and quarterly simulated *GNP* and flow of funds accounts, the operations of the actors within a smaller unit of time are represented. A sample "round" of activities which includes hiring and firing, wage payments and purchases of all sorts is represented as taking place 48 times each calendar year. Since the basic period is approximately a week, it is possible to keep track of unemployment durations of the simulated individuals almost in their customary units.

The forms of behavioral equations used in the model to control the simulated decision makers are derived largely from economics and business literature and the parameters have been adjusted to minimize the mean squared simulation errors of dynamic runs of the model in the period 1973–75. Money wage rates are set by the firms to change (weekly) at an annual rate of 3.0 percent plus .65 percent for every 1.0 percent change in the cost of living and by an additional −1.0 percent for every point the unemployment rate[8] is greater than 5.0 percent, subject to a weekly maximum equivalent to an annual rate of 21 percent per year. With the exception of the firms representing the farm, auto and mining industries whose prices have been set in all runs to follow the course they actually took, prices are raised by the firms when profit margins over average cost sink below averaged past profit margins. Firms figure the average cost for their current output based on labor cost, productivity, materials cost, depreciation, and interest and indirect taxes. Most of the wages paid to white collar workers, plus depreciation and interest are treated as fixed costs. As a result, average cost curves have a slight downward slope in the range of output in question. Although firms which run into production bottlenecks are programmed to raise prices on this account, this did not occur in the period in question. The labor force is programmed to rise by .25 persons for each additional regularly employed person in addition to a trended rise. Expectations of the future figure in the current version of the model chiefly in connection with decisions concerning real and financial investment; firms making wage and price decisions do not take forecasts into account.

III. Simulated Results

For the years 1973–75 the Transactions Model was run three different ways: 1) with federal fiscal outlays exogenously set equal to their actual values in the period; 2) with federal outlays augmented by a "low option" *PSE* program which allowed entry after 10 weeks of unemployment at a wage of 75 percent of private sector wages for each person's occupation and skill level to a maximum of $150 per week and 3) with a "high option" *PSE* program which allowed entry after six weeks of unemployment, at 75 percent of private sector wages up to a maximum of $225. Income taxes and social security taxes were collected on *PSE* wages.

Table 1 gives the results of the three simulations.[9] Neither of the *PSE* programs simulated was capable of preventing a rise in unemployment rates although the high option program succeeded in arresting the rise in unemployment rates by the end of 1974. Both programs significantly moderated the force of the recession on the labor market at a moderate net cost to the federal government.

Neither the low nor the high option program had a big effect on reducing unemploy-

[8]The *PSE* jobholders were counted as employed in calculating the unemployment rate for these purposes and in the table.

[9]Although we have shown actual figures for unemployment, the price index and *GNP*, so as to give an idea as to the ability of the model to track the data, the simulated run with no *PSE* program is best used as a baseline in assessing the simulated results of the *PSE* programs

TABLE 1—SIMULATED EFFECTS OF TWO VARIANTS OF A *PSE* PROGRAM
(All dollar figures in billions at annual rates)

	Civilian Unemployment Rate (%)				% of Civilian Labor Force on *PSE*		*GNP* Deflator			
	Actual	Simulated			Simulated		Actual	Simulated		
Date	No *PSE*	No *PSE*	Lo *PSE*	Hi *PSE*	Lo *PSE*	Hi *PSE*	No *PSE*	No *PSE*	Lo *PSE*	Hi *PSE*
1973 I	5.0	5.0	4.8	4.6	0.3	0.7	150	152	152	153
1973 II	4.9	4.6	4.6	4.5	0.3	0.7	153	154	154	155
1973 III	4.8	4.7	4.5	4.7	0.4	0.7	156	158	158	158
1973 IV	4.7	4.6	4.7	4.7	0.2	0.6	159	160	161	161
1974 I	5.1	4.9	4.7	4.6	0.4	1.0	164	164	165	164
1974 II	5.1	5.5	5.0	4.8	0.5	0.9	167	166	166	165
1974 III	5.5	6.1	5.6	5.4	0.6	0.8	172	170	170	169
1974 IV	6.6	7.0	6.5	5.8	0.4	1.2	178	174	174	173
1975 I	8.3	7.8	6.7	5.5	0.8	1.6	182	177	177	175
1975 II	8.9	8.0	6.4	5.0	0.9	1.6	184	181	181	179
1975 III	8.4	8.2	6.8	5.2	0.8	1.3	186	183	183	181
1975 IV	8.3	8.4	6.5	5.0	1.5	1.5	189	188	188	186
MEAN	6.3	6.2	5.6	5.0	0.6	1.1	170	169	169	168

	Gross National Product				Gross Cost of *PSE*		Net Increase in Federal Deficit		Net Increase in S/L Deficit	
	Actual	Simulated			Simulated		Simulated		Simulated	
Date	No *PSE*	No *PSE*	Lo *PSE*	Hi *PSE*	Lo *PSE*	Hi *PSE*	Lo *PSE*	Hi *PSE*	Lo *PSE*	Hi *PSE*
1973 I	$1,249	$1,260	$1,262	$1,264	$ 1.9	$ 4.7	$ 2.1	$ 4.8	$−0.6	$−1.7
1973 II	1,278	1,275	1,278	1,281	2.0	4.5	3.7	5.2	−0.7	−1.9
1973 III	1,309	1,300	1,302	1,305	2.6	4.7	3.0	5.2	−1.0	−1.5
1973 IV	1,344	1,322	1,325	1,329	1.5	4.1	3.2	4.9	−0.5	−1.8
1974 I	1,359	1,355	1,359	1,362	3.0	6.4	3.2	5.1	−1.5	−3.0
1974 II	1,384	1,371	1,376	1,376	3.7	5.7	3.5	5.9	−1.7	−2.7
1974 III	1,416	1,397	1,403	1,402	4.2	4.6	2.7	4.1	−2.3	−3.2
1974 IV	1,431	1,419	1,432	1,434	3.3	8.3	1.4	4.8	−2.5	−4.1
1975 I	1,417	1,444	1,471	1,482	5.6	11.5	−1.2	−0.9	−5.7	−10.4
1975 II	1,441	1,491	1,519	1,541	6.1	11.8	2.2	1.1	−5.9	−12.0
1975 III	1,498	1,530	1,555	1,581	6.2	10.2	4.6	−1.0	−4.7	−9.4
1975 IV	1,541	1,566	1,597	1,622	11.6	11.1	5.6	−2.4	−7.3	−9.7
MEAN	1,389	1,394	1,407	1,415	4.3	7.3	2.8	3.1	−2.9	−5.1

ment rates in 1973, because during that period relatively few people were eligible for it under the rules set up, and a high proportion of those who were eligible had entitlement to *UI* and did not go on the program. However, starting in 1974–II the situation changes markedly; both the number of people on either program and the direct expenditures on them are substantial through the end of the period simulated. During the period, the low and high option programs enrolled a maximum of 1.4 and 1.5 million persons, respectively.[10]

We have used the difference in the simulated federal deficit in a *PSE* program run and in the baseline simulations run as a measure of the net

[10]We have not addressed the issue of what this number of *PSE* participants would do. This number is considerably above the upper limit for job creation in the public sector set by Levitan in the National Commission volume (p. 165).

cost of that program to the Federal Treasury. Net costs are not very different from gross costs in the first half of the period, but in the second half the extra tax revenues generated through the moderating of the recession push net costs down significantly. Throughout the period the state and local governments gain through lower payroll costs and higher tax collections.

Neither form of the program causes simulated business firms to jack up prices any faster than prices rose in fact. Although the reduction of unemployment entails extra wage increases, their effect on unit cost is balanced by the effect of higher output levels and a greater stock of the newer and more efficient capital vintages. Whether one believes this is a realistic representation of what might have occurred depends on whether one believes that the model's decision-making rules for wages and other prices are reasonable. Replacing these rules by others would produce very different price results.

IV. Conclusion

Our simulation results would indicate that *PSE* programs can make a timely and significant reduction in unemployment in periods in which private demand is weak, at net costs to the Treasury which seem remarkably low when considering the benefits A *PSE* program in the form we have proposed would be a powerful addition to the automatic stabilizers we already have and would give direct help to those with the longest duration of unemployment. Consolidated net costs of these programs to all levels of government are modest, since avoiding the worst of the recession gives a boost to tax collections. The *PSE* programs simulated have the same fiscal effect as extending the eligibility for *UI* and increasing its benefits would have. The difference, of course, is in the services which the *PSE* job holders would perform. What the nature of those services might be and with what efficiency they might be performed are matters about which our calculations can say nothing. We believe that the opponents of the program are not wrong to worry about these matters, and that the proponents of the program should worry about them more. What our calculations do seem to show is that from the strictly macroeconomic point of view, programs of the type envisaged under the Humphrey-Hawkins framework deserve continued serious attention and research.

REFERENCES

Michael C. Barth, "The Full Employment and Balanced Growth Act of 1976: An Analysis and Evaluation," Institute for Research on Poverty Discussion Papers, Madison 1976.

Robert L. Bennett and Barbara R. Bergmann, *A Microsimulated Transactions Model of the U.S. Economy*, mimeo, 1976.

Barbara R. Bergmann, "A Microsimulation of the Macroeconomy with Explicitly Represented Money Flows," *Annals of Economic and Social Measurement*, 1974, 3/3, 475–89.

Alan Fechter, "Public Employment Programs," An Urban Institute Reprint, Washington 1976.

A. Gartner, W. Lynch, Jr., and F. Riessman, *A Full Employment Program for the 1970s*, New York 1976.

Charles L. Schultze, "The Economics of the Full Employment and Balanced Growth Act of 1976," Statement before the Senate Committee on Public Welfare, Subcommittee on Unemployment, Poverty and Migratory Labor, mimeo, Washington, May 14, 1976.

Lester Thurow, *Generating Inequality; Mechanisms of Distribution in the U.S. Economy*, New York 1975.

Michael Wiseman, "Public Employment as Fiscal Policy," *Brookings Papers*, 1976, *1*, 67–104, Washington 1976.

Congressional Budget Office, "An Economic Analysis of the Full Employment and Balanced Growth Act of 1976," May 21, 1976, Washington, DC.

National Commission for Manpower Policy, Proceedings of a Conference on Public Service Employment, May 1975, Washington, DC.

Simulation of Schumpeterian Competition

By RICHARD R. NELSON AND SIDNEY G. WINTER*

In modern formal theory, the virtues ascribed to competition are the virtues of an achieved state of efficient allocation, an essentially static condition that can be "sustained" by prices and price-taking calculations. Joseph Schumpeter's vision of competition, on the other hand, is a vision of an ongoing dynamic process, of a market system generating irreversible change in the course of historical time. The difference is profound; if it were not the case that both sets of ideas are presumably intended to illuminate the same reality, one would certainly regard them as belonging to separate subjects.

The key point is that in the modern competitive equilibrium story, what *can* be done is objectively and clearly defined. The question—both for the individual actor confronting his choice set, and in the analysis of the system as a whole—is what *should* be done. In the Schumpeterian scheme, the limits of what can be done are never fixed and never clearly in view. Discovering what can be done is part of the problem for the individual actor, and in analysis of the wider system, its performance as a social device for probing and expanding the limits of the possible is the fundamental concern. (Of course, the question of what *should* be done remains, and becomes more difficult as a consequence of the vagueness of the opportunities.)

Because of this central difference, a number of specific features of economic reality and specific theoretical approaches are seen in a very different light. Information imperfections, and informational differences among the actors, are not complications of the basic structure, but are central to the Schumpeterian scheme. The gains obtainable by guessing better and acting sooner are not a mere will-o-the-wisp, luring the actors toward inevitable frustration in equilibrium, but are the crucial motive power and adaptive mechanism of a system that is permanently in disequilibrium. And, because it arises from a continual unfolding of unanticipated possibilities, the disequilibrium is disequilibrium in the fundamental sense: Expectations are not being realized; mistakes are being made and corrections attempted.

It is plausible that the task of developing formal models is intrinsically more difficult in Schumpeterian theory than in modern orthodoxy. As the above remarks should make clear, the most powerful abstractions and simplifications of orthodoxy are inappropriate or ineffective in the Schumpeterian context. As we have argued elsewhere, the absence of formal theory probably accounts for the relative neglect of Schumpeterian ideas, ideas that, at the informal "appreciative" level, many economists find productive and persuasive.

Our purpose here is to discuss, and briefly illustrate, the role that simulation can play in the development of a formal theory of Schumpeterian competition.

I. Simulation as a Tool of Theoretical Inquiry

Simulation techniques have a wide variety of uses. The considerations relevant to the design and evaluation of simulation models are correspondingly diverse. In some cases, the designer of the model may have a high degree of confidence and clarity as to what model he wants to explore. His ignorance relates strictly to specific properties of the output, and his plan is to relieve that ignorance by using the computer to

*Professors of Economics, Yale University. We are indebted to Larry Spancake for research assistance. Financial support for development of the simulation model used herein was provided by the National Science Foundation under grant SOC-705529; this support is gratefully acknowledged.

generate the output logically implied by the input. In such a context, the advantage of simulation relative to an analytic approach is strictly a matter of feasibility or cost; if an analytic solution could be achieved by a comparable level of effort, it would be preferred to an accumulation of simulation runs. However, in many such cases it is impractical or impossible to obtain analytic results from the model the analyst has in mind. Rather than change assumptions and then analyze the wrong model, he may prefer to simulate the right one.

Our studies of Schumpeterian competition illustrate a different type of application of simulation techniques. We are uncertain, not merely about the specific implications of given assumptions, but about the assumptions we want to make and even about some of the questions we want to explore. In this sort of application, the relationship between simulation and analytic techniques is more complex than when the model and the questions are clearly specified at the outset. Freedom from the tractability constraints of available analytical techniques remains the major advantage of simulation; this freedom can be exploited in a preliminary exploration of a variety of alternative model formulations. Simulation also makes possible the sort of instruction of one's intuition that comes from careful scrutiny of detailed quantitative examples; this instruction is highly important when the models under study are stochastic processes, for such models often behave in ways that violate uninstructed intuition. Study of a range of examples may also lead to the perception of unanticipated regularities, and thus the development of new hypotheses about the behavior of the model.

Of course there are costs involved in working with a simulation rather than an analytic model. The most obvious of these is that the results are of uncertain generality. If there is large domain of interesting independent variables and parameters to explore, it is virtually impossible to explore all parts of it. This problem is compounded if the model is stochastic; one is then not sure about the representativeness of the results even for the parts of the domain explored. In addition, it is sometimes argued that the very freedom of assumption afforded by simulation modeling invites sloppiness, whereas the analytic approach imposes a need to make judgments as to which considerations are of central importance. In our view, the most serious problem with many simulation models, although the characteristic is shared by many analytic models, is lack of transparency The model yields results that are not easy to understand.

It would be a mistake, however, to pose the issue in terms of either-or. In a context set by specific questions about a specific model, the two approaches may be substitutes, but they are clearly complementary in the sort of exploratory effort that we have under way. We wish to stress, as a major theme of this paper, the extent of this complementarity and the size of the benefits potentially achievable if it is exploited effectively. Analysis is, on this view, an important component of a good simulation study. For example, it can yield predictions of the model's behavior in special cases, thus providing a check on whether the computer program actually functions in the manner intended. It can provide an interpretive framework for the results of simulation experiments, relating those results to the central ideas underlying the "black box" of the program. When unanticipated regularities appear in simulation results, it can be used to probe for an economically significant explanation. Simulation, on the other hand, can be a useful adjunct to an analytical approach It can establish, with the same finality as a theorem, the logical consistency of the model's assumptions with a set of propositions about its behavior. And while it offers a way around the tractability constraints of analytic methods, it imposes its own constructive discipline in the modeling of dynamic systems: the program must contain a complete specification of how the system state at $t + 1$ depends on that at t and exogenous factors, or it will not run.

The opportunity for fruitful exploitation of this complementarity can, however, be largely

foreclosed if it is not treated as an important consideration in the design of the simulation model. Most importantly, the freedom associated with the relaxation of tractability constraints must be exercised with restraint if the output is to be susceptible to analytic checking and interpretation. To introduce complexity in the name of "realism" alone, disregarding the added costs of checking and interpretation, is no more appropriate in the one theoretical endeavour than the other. It is, in short, a very pernicious doctrine that portrays simulation as a nontheoretical activity, in which the only guiding rule is to "copy" reality as closely as possible. If reality could be "copied" into a computer program, that approach might be productive—but it can't, and it isn't.

II. A Model

Our model of Schumpeterian competition involves an industry composed of firms interaction in an environment characterized by an exogenously changing fund of research opportunities, an endogenous output price, and a number of cost, technical and decision rule parameters. Firms follow long-term policies for acquisition of new technical information: they may do "research," which is an attempt to exploit the exogenously changing research opportunities; or they may do "imitation," an attempt to acquire new information by copying the practices of other firms; or they may do both. Research and imitation are both costly, and in both cases a firm's policy is expressed as a commitment to a certain rate of expenditure per unit capital, so information acquisition efforts change proportionally as a firm grows or declines.

The richness of the fund of research opportunities is characterized in any period by a single number, a level of output per unit capital that we call "latent productivity." Latent productivity represents the central tendency of a distribution of research outcomes Research expenditure buys, in each period, a probability of a draw on the distribution of research outcomes; the probability is proportional to the expenditure, and parameters are set so that the upper bound of probability one is not encountered. Imitation expenditure buys a probability of a draw from a distribution determined by the current productivity levels of other firms. When a firm obtains, by research or imitation, a technique superior to its current one, its entire capital stock is switched (costlessly) to the new productivity level in the following period.

Firms utilize all of their capital in each time period, and thus produce an output equal to their capital stock times their current productivity level. An industry demand curve determines output price from the combined outputs of all firms. Firms' desired investment is determined by the relation between price and marked-up unit production cost; desired mark-up is an increasing function of the firm's market share—and thus a profitable firm exercises restraint in expansion when it is large relative to the market.

III. Concentration as Cause and Effect of Progressiveness

One of the major themes associated with Schumpeter is the so-called "Schumpeterian hypothesis"—the claim that the static allocation disadvantages associated with market power may be more than offset by the greater technical progressiveness of larger firms. Schumpeter's original discussion in *Capitalism, Socialism and Democracy* wove a number of quite distinct arguments around this theme. In our recent work, studying the problem in the explicit dynamic setting of the simulation model described above, additional conceptual issues have been identified, and ones previously identified have been highlighted. For the purposes of this brief report, we will focus on one part of the complex total picture, namely, on the "reverse-Schumpeterian" linkages through which market structure is itself affected by the conditions of technical progressiveness in an industry.

We abstract here from a number of familiar determinants of market structure that are of obvious importance in some real situations, in-

cluding barriers to entry and economies of scale in production. We are concerned with how concentration develops as a consequence of an accumulation of chance differences in firm success, in a context in which firm decision rules and capital market considerations tend to make growth rates approximately independent of firm size. This is conceptual territory occupied by the stochastic theories of the firm size distribution, in which the contributions of Herbert Simon and his collaborators are particularly noteworthy. (See Y. Ijiri and Simon.) These theories are rather narrowly focused on the implications of stochastic growth rate differentials for the shape of the firm size distribution; they do not explore the economic sources of the growth rate differentials, or relate firm growth to investment decisions, or detail the environment in which the firms interact. However, a number of authors have noted, and Almarin Phillips has particularly stressed, that industries in which rapid technical change is going on are likely to be ones with large variances in firm growth rates. It has been argued, informally, that this creates a link between concentration and progressiveness with the reverse causal sense from the Schumpeterian hypothesis. Industries with high technological opportunity are likely to have high realized progressiveness, and, since research is an uncertain business, they are also likely to have high interfirm growth differentials and thus develop, over a period of time, high concentration.

IV. Some Illustrative Results

Our model serves well to formalize, illustrate and elaborate this informally argued hypothesis. In it, the uncertainties associated with technical change provide the sole stochastic element. We may associate the notion of "greater technological opportunity" with a higher rate of growth of latent productivity. And the model is linked to the theories of the firm size distribution by a modified Gibrat's Law assumption: other things equal, two firms of different size (capital stocks) will differ in gross investment in proportion to their size, except to the extent that the larger firm perceives itself as having greater market power.

In a forthcoming paper we report the results of a simulation experiment involving initial industry structures of 2^k equal-sized firms, $k = 1, \ldots, 5$, in a particular regime of growth of latent productivity. In each case, half of the firms were "researchers," and did both research and initiation, and half were "imitators" and did only imitation. Only research was significantly costly. The rate of latent productivity increase was set at 1 percent per quarter, and the demand curve facing the industry was fixed and of unitary elasticity. Each run simulated the industry's development over 100 quarterly periods.

One very clear result of the experiment was the tendency of concentration to increase substantially over the run in cases where there were 8 or more firms initially. For a descriptive measure of concentration, we used the "entropy numbers equivalent"—the number of firms in an industry of equal-sized firms that would give the same measured entropy as the observed distribution. In the 32 firm cases, the entropy numbers equivalent at the end of the run averaged less than 10. The increase in concentration was comparable to the outright elimination of over two-thirds of the firms initially in the industry if the survivors had been of equal size.

Recently we have been engaged in further explorations of the development of concentration and its relationship to progressiveness and the relative success of researchers and imitators. One part of this inquiry involved a replication of the earlier experiment, except that the rate of latent productivity increase was doubled. The impact on concentration at the end of the run is shown by comparison with the result of the earlier experiment in Figure 1. The qualitative result is clear. There is very little effect when concentration is initially high (and unchanging in the course of the run). However, the Phillips hypothesis is confirmed strikingly when concentration is initially low (and increasing throughout the run). It is interesting that the 32 firm runs actually end up at a slightly *higher*

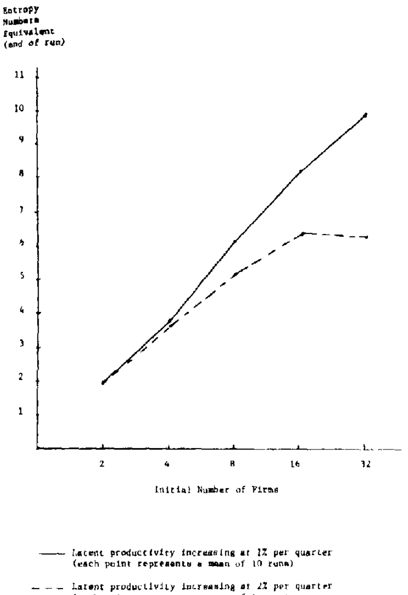


FIGURE 1

level of concentration (smaller numbers equivalent) than the 16 firm runs, when latent productivity advances rapidly! This may well be a result of sampling variability, but it is perhaps worth noting that there is nothing logically wrong with such a result. Two size distributions that give the same entropy numbers equivalent (or agree on any other numerical descriptor) may have quite different stability properties. When the numbers equivalent reaches 16 in a 32 firm run, the size distribution is quite different from the initial distribution of a 16 firm run.

The impact of the higher rate of latent productivity increases on the development of concentration, in the context of our model, may be explained as follows. Imagine that firm sizes are artificially held fixed. Then, the implications of constant research and imitation policies could be described in terms of a characteristic distribution of *time lags* between latent and achieved productivity—e.g., a firm of a particular size may tend to be on the average four periods behind latent productivity. The larger the rate of latent productivity increase, the greater the productivity difference associated with a given lag, and hence the larger the profitability stakes involved—and growth rates depend on profitability. Thus, the variance of the growth rate is greater when the rate of productivity increase is higher.

But an obvious question arises as to why concentration is not much affected when it is initially high. We have investigated this question with a combination of analytics and additional simulation runs, but the conclusions here expressed are tentative. One major consideration is the investment and hence output restraint exercised by firms that acquire substantial market shares. Under our assumptions, when there are two large firms sharing markets relatively equally, and one gains a productivity advantage, it will not exploit this aggressively for fear of spoiling the market. Thus, the temporarily disadvantaged firm will not be forced to shrink rapidly in size and research effort, and can subsequently make a comeback. We manipulated this restraint experimentally by changing the target markup formula; essentially, we imposed on our firms a subjective belief that industry demand is much more elastic than it actually is. This led to some 2-firm runs with a final numbers equivalent of 1.5 or below—but to our surprise, it also produced additional runs with numbers equivalents of 1.9 or above. Clearly, investment policies were not the only explanation for the stability of structures involving a small number of equal-size firms.

A second major consideration is the lower sampling variability in individual firm productivity levels associated with a higher rate of research and imitation draws. Under our assumptions, a larger firm has (for a given policy) a proportionally larger probability of getting a draw in any period. This means not only that the achieved productivity level is higher on the average, but the variance of that productivity level is reduced; consider, for intuitive justifica-

tion, the case where there is no research outcome variance around latent productivity and the probability of a draw approaches one. Add to this the fact that it is really the dispersion of results within a *sample* of research outcomes that is relevant; there is no differential growth if all firms chance to succeed or fail together.

For a concentrated industry to display the same interfirm variability in productivity as a fragmented one, the research environment of the concentrated industry must be more uncertain, or imitation must be relatively impeded. In the model, the uncertainty in the research environment can be increased by increasing the dispersion of research outcomes around latent productivity, or by reducing the frequency of research draws. The tendency of imitation to keep the industry clustered in productivity levels can, of course, be adjusted by changing the assumed imitation policies. As a check on our understanding of the mechanisms at work, we have done some runs with assumptions that imply greater research uncertainty and less imitation; the results tend to confirm the expectation that even an industry initially structured as a few equal-size large firms can display sharply increasing concentration in such a technological environment.

Schumpeter, in his verbal theorizing, pointed out long ago the "routinization of innovation" that occurs in industries with a few large firms. Our model formalizes the logic of one mechanism by which such greater determinateness can come about. Its implication is that industries with a few large firms are much less subject to stochastic drift to still higher concentration than are fragmented industries, and the pace of the drift is relatively much less responsive to the rate of advance of technological opportunities.

REFERENCES

Y. Ijiri and Herbert A. Simon, Interpretations of Departures from the Pareto Curve Firm-Size Distributions," *J. Polit. Econ.*, Mar.-Apr. 1974, *82*, 315–31.

Richard R. Nelson and Sidney G. Winter, "Neoclassical vs. Evolutionary Theories of Economic Growth: Critique and Prospectus," *The Economic Journal*, Dec 1974, 84, 886–905.

——— **and** ———, "Dynamic Competition and Technical Progress," forthcoming in *Economic Progress, Private Values and Public Policy: Essays in Honor of William Fellner*, B. Balassa and R Nelson, eds.

Almarin Phillips, "A Study of the Aircraft Industry," *Technology and Market Structure*, Lexington, 1971.

Joseph A. Schumpeter, *The Theory of Economic Development*, Cambridge, Mass. 1955.

———, *Capitalism, Socialism and Democracy*, 3rd ed , New York 1950.

Competition and Market Processes in a Simulation Model of the Swedish Economy

By Gunnar Eliasson*

This paper presents the outline of a modeling project on the Swedish economy that was started 1-½ years ago.[1] Some special features are highlighted (verbally and diagrammatically) and the paper concludes with a numerical (*NB! not* empirical) illustration from the Swedish labor market. The purpose of this paper is not to present the mathematical specification needed for a full understanding of the mechanics of the model system.[2]

The essential idea of the model is to integrate micro activities through markets in order to get a better grasp of the behavior of conventional national accounts aggregates. We can also formulate the purpose to be to systematize micro information for a richer understanding of economic behavior at the national level. We have chosen to identify the micro unit with a financial decision unit (the firm with controlled subsidiaries) and a household.

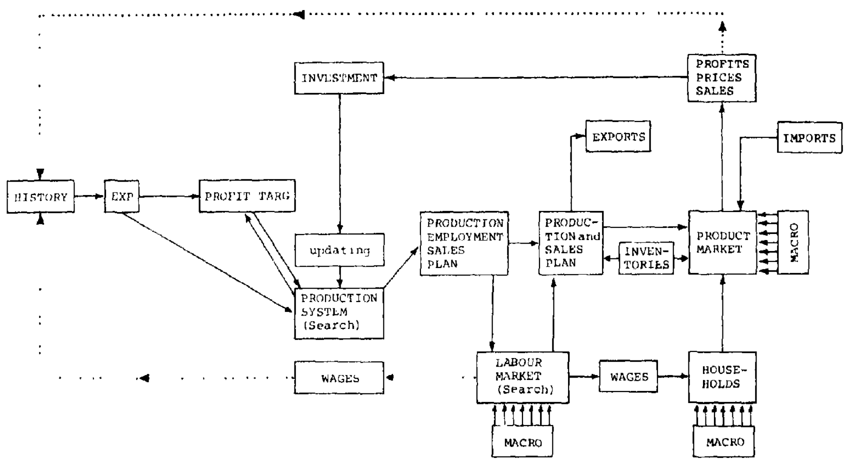


Figure 1 Business Decision System (One Firm)

So far the model is only specified at the micro level for the business sector. Figure 1 indicates how one firm interacts with all other firms and with the macro model through labor and product markets. It should therefore be called a micro industry model contained within a (national accounts) macro framework.

Besides the explicit combination of micro and macro levels through imitation of the market process, two, (somewhat novel) features are highlighted in this paper:

1) the determination of the firm's profit targets and the firm's responses to these targets (internal market pressure);

2) the specification of the search for satisfactory labor output combinations *within* the

*Chief Economist, Federation of Swedish Industries. Gosta Olavi and Mats Heiman at *IBM* Sweden have been extremely helpful in getting the model consistently specified and programmed

[1]As a joint project between *IBM* Sweden and the University of Uppsala

[2]Interested readers are referred to Eliasson (1976b)

firm's production system.

Priority has systematically been placed on realistic specification at the expense of simple, conventional parameter estimation. Until now specification work has proceeded as if no such problems existed, except that all variables entering the model have a known and measurable counterpart. Hence, simulation experiments will be necessary to calibrate many of the parameters of the model. This calibration phase has begun recently on the basis of a skeleton model.

Conventional econometric methods will be used when possible to estimate parameters, especially of the macro model. The limited space does not allow an account of our methods on this point (Eliasson 1976b). Until we have found a numerical specification of the model to believe in, the experiments performed on the model should be regarded as theoretical (albeit numerical) analysis. We will present and interpret an experiment at the end to help explain how the model works, not to draw conclusions about the Swedish economy.

I. The Model

The skeleton model that we are now working on is probably best explained by Figure 1 on the business decision system (the micro unit). In the *EXP* block a vector of historic wage, price, and sales, etc. data are transformed *each quarter* into expectations. Conventional smoothing formulae are used here and in the *TARG* sector, where past profit performance is transformed into targets or profit goals (see below).

Expectations and imposed targets start up a search sequence within the production system of each firm each *quarter*. Search concludes with a quarterly production and recruitment plan that satisfies profit *TARG* requirements.

Firms then lay off redundant labor or search the labor market for additional labor (see below). When the labor force has been fixed, the next quarter's production plan and wage bill are also given. Some products are marketed abroad at exogenously given foreign prices. The export share changes in response to the relative movement of foreign and domestic prices. The rest feed into the domestic markets together with imports. Wages and salaries are added up and transformed into money demand by a macro Stone type expenditure system (bottom right hand corner, Figure 1). As the model stands now, firms communicate an offering price (based on their expectations) to households, who respond by telling how much they will buy. If this is more than firms have *already* decided to supply, firms step up their offering price and vice versa. This goes on 3 times. Then prices and purchase volumes are assumed given, and inventory change is distributed over firms in a rough and ready way.

Prices, wages and output having been determined, each firm can calculate its quarterly profits, and data are fed back to update expectations and targets In the process a provisional cash flow model determines investment to update the production system.

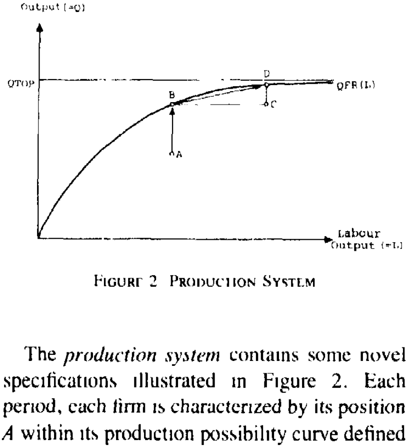


FIGURE 2 PRODUCTION SYSTEM

The *production system* contains some novel specifications illustrated in Figure 2. Each period, each firm is characterized by its position *A* within its production possibility curve defined as a function of labor input (*QFR* in Figure 2). The vertical distance *AB* measures how much output can be increased without hiring any new labor and *CD* the extra increase that can be obtained by the application of additional labor.[3]

[3]Since 1975 estimates on *AB* and *CD* are asked in the annual planning survey of the Federation of Swedish Industries. The coordinates of these points allow us to approximate the production possibility frontier *QFR*

The functional form of *QFR* is:

$$(1) \qquad Q = QTOP \cdot (1 - e^{-\gamma L})$$

which has the desirable convexity properties. Marginal labor productivity furthermore equals $QTOP \cdot \gamma$ in origin. There is no explicit capital stock or capital service measure in the model.

Productivity of new investment is entered exogenously. The amount invested is determined endogenously through a profit-plow-back mechanism. These two factors together shift the asympole *QTOP* upwards and bend *QFR* through changes in γ simultaneously.

The preliminary sales expectation initiates a search sequence, the exact nature of which depends on whether sales expectations mean recruitment or not and on the strictness with which targets are enforced. It is important to note here that this makes realized average labor productivity change endogenous. *QFR* is updated by investment each period, and slack is both explicit and endogenous in the model. This specification allows firms under internal profit pressure occasionally to upgrade their productivity by moving somewhat beyond *QFR*.

II. Markets and Competition

We will deal with two aspects of markets and competition in this paper, namely:

1) the internally imposed management pressure to perform in response to realized profit experience (targeting);

2) labor market search.

The *labor market* is at present too simplified. Labor is homogeneous, and a firm can search the entire market and raid all other firms subject only to the constraint that search takes time. We have solved this provisionally by having firms scan the market for additional labor randomly, the probability of hitting a source being proportional to the size of the firm (number of employed) and the size of the pool of unemployed. Six trials (iterations) are allowed each period. A firm-to-firm confrontation means an upgrading of the relatively low wage firm with a fraction (.20) of the wage differential. A firm being searched can lose a maximum of one percent of its employees, having to upgrade its wage level each time and being forced to reconsider its production plan. We will study this relatively competitive labor market with and without a device (a regulation) aimed at preventing layoffs without ample advance notice. And we know already from experimentation that small modifications in the labor search-wage response coefficients affect macroeconomic behavior strongly.

Firm *targeting* is decisive for the behavior of the entire industry sector. The device as well as the systematic concentration on operating profit margins at the firm's headquarter can be backed by empirical evidence (Eliasson 1976a). The *feedback* historical reference target is defined:

$$(2) \qquad MHIST(t) = \lambda \cdot MHIST(t-1) + (1-\lambda) \cdot M(t-1)$$
$$0 \leq \lambda \leq 1$$

On this we apply the targeting principle *maintain or increase* (MIP) profit performance:

$$(3) \qquad TARG(M) = MHIST \cdot (1 + \epsilon)$$

$$(4) \qquad \epsilon \geq 0, \text{ but small.}$$

M stands for operating profits in percent of sales. Expressed in words, a historic "reference" profit level is updated each period. This reference, raised by the fraction ε is then used as a profit requirement when the firm makes up its plans.

All solutions (production plans) are based on expected prices and wages and on the coefficients of the production system. Plans are checked each quarter against *TARG(M)* before being fixed.

TARG(M), as in (3), is calculated once a year. Management response to current market performance of the firm can be tuned by varying λ, ε and the persistency of enforcing annual targets through the quarters of the year.

III. State of Numerical Specification

Empirical knowledge enters the model in 7 ways: 1) the causal (hierarchial) ordering of

model "modules," 2) structural parameters, 3) time reaction parameters, 4) start-up positional data, 5) start-up vector of historic input-data, 6) parameters of macro-model and national accounts identities and 7) exogenous inputs (productivity level for new investment goods and foreign prices).

(1) is highly important for the behavior of the entire system. Here we believe to have a satisfactory empirical specification already, based on Eliasson (1976a). For the time being (2), (4) and (5) have been obtained by randomizing macro data for each of five subindustries for 1970 through 1974 into individual firms. Thus the historic input vector begins relatively low and winds up with the inflationary peaking of the business cycle in 1974.

As for (6) we use macro estimates from elsewhere, mainly the household sector.[4]

Equation (3) constitutes the critical area where essential empirical knowledge is missing. Time reaction parameters are found mainly in the *EXP*, *TARG* sectors and *SEARCH* algorithms in the production system and the labor market. (Note that lack of empirical knowledge here refers to the speed of response, not to the causal ordering of decisions and *how* the firm reacts.) No empirical research on the format of this model is available on this point. The procedure will have to be to experiment with alternative specifications and to assess results by comparing them with observed behavior in official macro statistics.

Experience so far is that we do not get stuck with a whole lot of specification alternatives that behave well and that we cannot discriminate between. We believe this to be the natural consequence of the constraints imposed already by the empirical knowledge brought in.

We have so far found difficulties getting the economy to grow and to behave cyclically without feeding it constantly with an exogenous growth cycle. This may be a relevant property of an internationally integrated economy like Sweden's, although I doubt it. Whatever one believes, this will be one of the prime questions to be investigated by further experimentation and by loading the model with real, individual firm data.

IV. A Labor Market Experiment

In the numerical (NOT empirical) illustration that follows we show only the relative economic development in index form with and without a labor market turnover restriction.

This device[5] requires that firms give most of employed labor 6 months advance notice before anyone can be laid off. In the model, redundant labor in individual firms is placed in two-quarter files. If redundancy persists, people in the second quarter file can be fired in the next quarter.

The device can be switched on and off in experiments. We have run a number of such experiments and the relative differences over time are systematic in all experiments.

In the cases tried *without* the device, the economy converges slowly towards a steady state growth path with a somewhat reduced rate of growth, that is not sufficient to keep the economy fully employed without an exogenous growth in public employment of about the same magnitude as during the last 10 years.

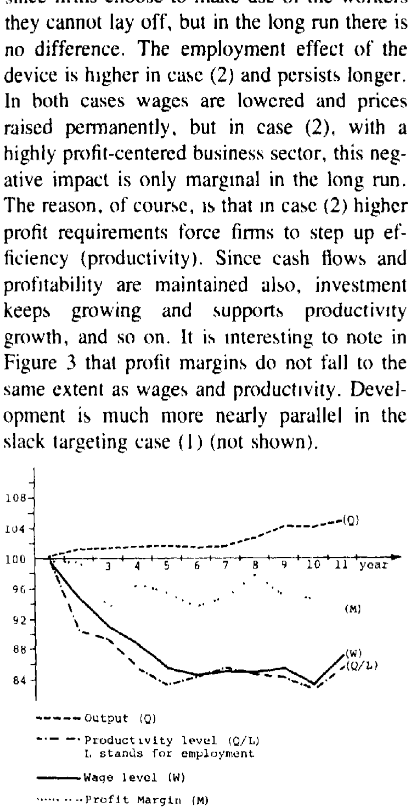
In a pair of experiments we introduced the device under (1) rather slack targeting requirements ($\lambda = 0.65$, $\varepsilon = 0$ in all firms) and (2) in a well-tuned, profit-centered business sector ($\lambda = 0.9$, $\varepsilon = .01$). Figure 3 illustrates case (2). Note first that without the labor market restriction case (2) is only marginally better at the macro level in terms of production growth, productivity, etc., than case (1). Furthermore, experiments indicate that the search process (number of iterations) is important for a cycle to emerge. Third, exogenous variables have been given average growth rates that correspond to their development in Sweden during the last 10-year period.

In the model the device operates as a permanent impediment to the attainment of business targets. In both cases profit margins suffer sub-

[4]See C. J. Dahlman and A. Klevmarken, who have estimated a linear expenditure consumption system for Sweden.

[5]That roughly corresponds to a law recently enacted in Sweden

stantially, but more so in the slack targeting case, where deteriorating performance spins off a gradual lowering of profit ambitions and also a gradual slowdown of growth. In the beginning the employment effect is positive and high; then it tends to fade away because of the slowdown in growth. This is probably not realistic. When target requirements are stiffer as in case (2), profit margins pick up again after a while. Growth is actually higher during the first years, since firms choose to make use of the workers they cannot lay off, but in the long run there is no difference. The employment effect of the device is higher in case (2) and persists longer. In both cases wages are lowered and prices raised permanently, but in case (2), with a highly profit-centered business sector, this negative impact is only marginal in the long run. The reason, of course, is that in case (2) higher profit requirements force firms to step up efficiency (productivity). Since cash flows and profitability are maintained also, investment keeps growing and supports productivity growth, and so on. It is interesting to note in Figure 3 that profit margins do not fall to the same extent as wages and productivity. Development is much more nearly parallel in the slack targeting case (1) (not shown).



Graphs show index (annual average) of simulation with labour market restriction divided by NO restriction case, under high target pressure (case 2)

FIGURE 3. LABOUR MARKET EXPERIMENT

Also observe the close correspondence between wage and productivity development that we have observed in all experiments run so far. The reason lies in the profit targeting, production search process, and variations occur when inflation rates differ in the long run.

The consequences of the introduction of the device probably depend on the initial positioning of the economy. If enforced in the beginning of a business upswing, it may mean very little. If introduced at the end of the upswing, just before growth comes down (as in our experiments), profits are immediately depressed (compared to the case with no restrictive device) and productivity growth comes down relatively since firms cannot lay off as many people as they desire, retirement from the labor force is unaffected, and new labor demand from other firms is low.

I have emphasized that we should draw no conclusions about Sweden from the numerical results. This may seem to reduce the reported experiments to a theoretical game, which is perhaps less interesting. I should modify this statement to the extent that much empirical information is already embedded in the model. Experiments mean constant confrontations with observation. This is part of the validation process that should increase our knowledge also of the parameter specification of the economy.

Personally, I would argue that the direction of responses we have studied is what should be expected on the basis of knowledge of both the model and the economy, although the model tends to overreact. Furthermore, calibration is only in its early stage. More tests are needed before any stronger empirical positions should be taken.

REFERENCES

C. J. Dahlman and A. Klevmarken, *Den privata konsumtionen 1931-1975*, Uppsala 1971.

Gunnar Eliasson, *Business Economic Planning, Theory, Practice and Comparison*, London 1976a.

———, *A Micro-Macro Interactive Simulation Model of the Swedish Economy*, Model specification, Federation of Swedish Industries. Preliminary mimeographed paper, Stockholm, Dec. 1976b.

BRITISH CAPITAL IN THE LATE NINETEENTH CENTURY: SOURCES IN BRITAIN AND MOVEMENT IN THE EMPIRE

Public Expenditure and Private Profit: Budgetary Decision in the British Empire, 1860–1912

By LANCE E. DAVIS AND ROBERT A. HUTTENBACK*

To imperial enthusiasts, Empire connoted the triumph of British principles over the powers of darkness and was a source of incalculable psychic and financial reward. The critics, however, saw an increasing burden on the British domestic taxpayer, while British subjects in the colonies contributed virtually nothing.

Any analysis of the cost of empire in the context of nineteenth-century Britain must in large part rest on an understanding of the institutional mechanism that might have provided the base either for exploitation or selfless regeneration of barbarous places.

In an earlier period, monopolies enforced by the military power of the state provided one such institutional structure. Later similar transfers might have been effected through the assertion of ownership over valuable resources in relatively fixed supply. However, even if British entrepreneurs had been omniscient enough to recognize such resources before their competitors, the profit opportunities must have been relatively limited.

By 1860 Britain was committed to free trade and the Empire was as a consequence theoretically open to all. Thus, possibilities for direct monopolistic profits were very small; and similar competitive forces acted to reduce monopsonistic rents. Given such an environment, any exploitation must have rested on a set of government policies that thwarted competition and gave some shadowy "imperialist" an edge over his colonial, foreign and even domestic rivals. Thus the degree to which empire was exploitative as opposed to burdensome cannot be determined without reference to the government sector—and that appraisal is the focus of this paper.

Every government policy involves a budgetary dual. Wars cannot be fought without armies being paid; tariffs can't protect local enterprise without some expenditure on enforcement, and even property rights cannot be guaranteed unless funds are devoted to legal and judicial needs. Of course, budgets are not always what they seem, but in the nineteenth and early twentieth centuries they did in large measure reflect the policies of government.

I. The Data and the Model

A study of the budgets of British colonies allows us to make certain inferences about government behavior, but a few words of caution on the "underdeveloped countries" are in order. The series are not always complete and there is some year-to-year fluctuation in national composition. International comparisons are only made at the level of the central government which, given the varying role of local governments, may present some problems. For the United Kingdom and the British colonies, however, we have compiled a series of total (all level) receipts and expenditures for every fifth year in the period 1860–1912. Finally, because of the many different ways of financing railroad development, all government receipts and ex-

*California Institute of Technology

penditures for this sector have been deleted.

This paper is primarily an exercise in regression analysis. We view the results more in the spirit of prescience, but it still appears to yield interesting insights into the operation of the imperial machinery.

The continuous independent variables include time and (time)^2, railroad mileage per square mile (an attempt to provide some index of development), and lagged values of the deviation of exports from a time regression (an effort to get at short run phenomena that may have affected the budgetary process). In addition, the world has been divided into eleven regions and six government groups. For the period 1880–1912, a population density variable (percent of population in cities over 3,000) has also been used for the imperial data.

The dependent variables are certain revenue and expenditure categories per capita and the proportion of the budget devoted to these same categories. If R^2's are chosen as a measure of success, the results are poor from an economist's standpoint and varied from that of an economic historian. They range from .75 to something less than .05. We, however, are more interested in the signs of the coefficients and these, we feel, are frequently of substantial interest

II. The Results

We begin with a series of conjectures about the behavior of an "ideal" imperial system designed to provide an efficient mechanism for the transfer of income from the colonies to the mother country, given the constraints previously enumerated. We focus on six classes of expenditures: 1) total, 2) administrative, 3) public goods, 4) deadweight from debt, 5) justice, and 6) defense.

In the case of total expenditures, theory provides only a tenuous basis for useful conjecture. If expenditures had been higher in the colonies than in the United Kingdom or in the developed independent world, this fact would be surprising, and if these high levels of expenditure included substantial remissions to the home country, we might have sufficient evidence to accept out of hand the asserted profitability of empire. Unfortunately, and not too surprisingly, such was not the case.

In the area of administration, it was at one time contended that the Empire represented nothing so much as a vast system of outdoor relief for the British upper classes. To the extent that this general charge is correct, one would expect per capita expenditures on administration to be higher in the colonies than in other underdeveloped areas and the percentage of budgets devoted to administrative expenses to be greater than in all other administrative entities under consideration. Furthermore, such expenditures should decline as a colony gained greater autonomy.

In terms of total public goods expenditures, the question is ambiguous. To the extent that such expenditures made trade and commerce easier and more lucrative and were paid for by the local residents, they should have been encouraged by the imperial authorities. Conversely those authorities should have been less concerned with those aspects of social overhead capital (education, for instance) that might have tended to promote domestic entrepreneurship and produce greater competition with the mother country As colonies attained increasing control of their own destinies, one might expect them to have altered their basket of public commodities in the latter direction.

If Britain had really been an imperial power whose only goal were the exploitation of its colonies, one would expect there to have been high deadweight debt expenditures. Capital markets have always been among the least competitive (at least in the spatial dimension) and it should have been easy to force the colonies to borrow in the London money market and so provide more income to British capitalists. Of course, one would expect debt charges to decline with increasing self-government.

It could reasonably be assumed that the easiest area for income transfer was in the field of defense expenditures. The transfer would not be direct, but by relieving the metropole of the re-

sponsibility for much of imperial defense, it would permit lower levels of public expenditure and/or investment in terms more directly productive. Once again expenditures should decline as autonomy increased, but dependent colonies should have spent a substantially higher proportion of their budget on defense when compared with the United Kingdom or with developed independent countries.

When we turn to the data we find that while the evidence indicates some conformity with our conjectures, the match is not so close as to permit us to accept the argument that the British were consciously using indirect government power as a mechanism for inter-regional income redistribution.

In terms of total expenditures, the costs of government do not seem to have varied either with spatial location or with form of government.[1] Over time, expenditures everywhere tended to fall, but at a decreasing rate. If we limit our scrutiny to the Empire but include all levels of expenditure, we find that expenditures per capita were lower in Type 2 and 3 colonies but higher in those with responsible government than they were in Britain. When the comparison is limited to central government expenditures and the area extended to include the whole world, the colonies as a group spent somewhat less than the United Kingdom (but only because the self-governing ones spent much more), the developed countries and the underdeveloped noncolonial world, substantially more.

In the case of expenditures on administration, the argument for direct exploitation becomes even weaker. While by European standards the British spent less money and a smaller proportion of their budget on administration and to a certain extent exported that taste for administration to their Empire, the results are quite different from what one might have expected. The colonies, taken as a group, spent less per capita on administration than did Britain; the noncolonial undeveloped countries spent more. The relation with the development variable is interesting because both within the Empire and in comparison with the rest of the world there seems to be a negative relation with per capita expenses on administration.

Over time within the Empire there is an increasing level of administrative expenditures but the rate of that increase declines steadily. Under no conditions does there appear to be much support for the view that over time there was a natural pressure to increase the level of administrative expenses. That conclusion is further supported by an examination of the percentage distributions of the budgets. Both for the world in general and for the Empire the $(\text{time})^2$ term is strongly negative. Taking all the colonies together they spent about the same proportion on administration as did the British. But on closer examination it appears that the percentage of the budget devoted to administration in the self-governing colonies was slightly less than in the United Kingdom, while the proportion spent in Types 2 and 3 was marginally higher, and those were the colonies that tended to employ British as opposed to local administrators. Still the budgetary commitment even in Type 3 colonies was far below the level attained in independent states.

An examination of the expenditures on public goods does appear to provide some support for those critics of empire who saw the whole enterprise as being based on exploitation. However, it is not clear that a more rapid rate of capital accumulation, while certainly working to the benefit of British business interest, did not also benefit the colonies. Independent countries spent more per capita than did either Britain or her colonies. Taking the world as a whole, the time trend is negative but the squared term significantly positive. Within the Empire the expanded budgets indicate similar time trends and show that urbanization does increase the level of per capita expenditures on public goods. All types of colonies spent more than the United Kingdom; however, self-governing colonies spent far more while for Types 2 and 3 the difference, though positive, is only marginally significant.

[1]Type 1. Self-government, Type 2. Significant local representation, Type 3. Dependent status.

A more interesting result is produced by a breakdown of public goods expenditures roughly into traditional and human capital. The British did not spend much on the latter category in Britain itself and it is not unreasonable to assume that they exported their tastes overseas. Not unexpectedly, the self-governing colonies spent substantially more than the United Kingdom, but that colonies of Types 2 and 3 spent more than Britain (although substantially less than the remainder of the developed world) is worthy of note. Even more surprising, however, is the fact that noncolonial underdeveloped states appear to have spent less on human capital than even Type 3 colonies.

The percentage breakdowns indicate similar trends, although they do perhaps suggest a slightly different focus. Within the Empire all classes of colonies spent a greater proportion of their budgets on public goods. When the comparison is made with the rest of the world there are relatively small differences between the developed nations, but all underdeveloped countries tended to spend more However, colonies of all types spent a much greater percentage of their budgets on public goods than did the independent nations. Only in India and in South Africa is a smaller percentage devoted to public goods. Once again, although traditional capital draws the bulk of the resources, the colonies fare better than either the United Kingdom or the independent underdeveloped countries in terms of their relative commitment to human capital

Without doubt the greatest spenders in both per capita and percentage terms were the self-governing colonies. It is clear that these investments were choices freely made. On the other hand, Britain was not ungenerous in the dependent colonies nor did she divert all the resources into those activities that would have provided the greatest rewards for the British businessman.

In the case of justice the results are more ambiguous. Expenditures in this area serve not only to maintain property rights but civil order and the legal contractual base for business. An examination of the regional breakdown indicates that South Africa, the Caribbean and the Mediterranean regions spent significantly more than the United Kingdom, but that India spent less. The first three cases may well indicate that the British used governmental power to maintain external conditions conducive to British (or for that matter anyone else's) business, but the last instance almost certainly reflects the expenditure complementarity that was discussed earlier. At all levels of government, the colonies as a whole spent more than Britain, but the level of their expenditure declined over time. Moreover, they spent substantially less than foreign countries. Within the Empire, the expenditure pattern appears slightly paradoxical. The largest expenditures were in self-governing colonies, while those with dependent status spent only marginally more than the *U.K.*

Turning from the per capita figures to the relative ones, within the Empire cities have the expected positive effect. However, while all classes of colonies spent a higher proportion of their budgets on justice than did the United Kingdom, it is the Type 2 and 3 colonies that spent a very much greater fraction. Furthermore, in the worldwide comparison, the proportion expended in the other developed countries was higher than in Britain, while the colonies (now more narrowly defined) still spent more than the mother country, and the foreign noncolonial nations about as high a proportion as the self-governing colonies.

One would expect British colonies to have been forced to devote a significantly greater proportion of their resources to justice than either Britain or other nations, and Type 2 and 3 colonies to have designated a greater percentage of their budgets to this subject than their self-governing fellows. While these conclusions are not true for absolute expenditures, they do hold when the measure is a relative one.

Debt deadweight should be examined in the light of revenues that are raised through loans. Colonies, developed and underdeveloped countries, all borrowed more than the United Kingdom but the undeveloped world borrowed slightly more than the colonies, and the developed much more. In part that latter difference

may reflect only the stage of development and in part it may only be a statistical artifact reflecting the failure to include subcentral units in the accounting scheme. Still it is clear that while colonies borrowed more per capita than Britain, the foreign underdeveloped nations borrowed even greater amounts. Loans tended to fall over time, but at a decreasing rate, and they are positively correlated with development. Within the Empire the inclusion of subcentral levels of government changes the results somewhat. The self-governing colonies borrowed much more heavily than the mother country, while colonies of Type 2 and 3 operated at about the same level. But given the lower incomes of those colonies, their borrowing rates were very impressive. When one turns to relative proportions, the conclusions are only slightly altered. Australasia, South Africa, North America and India were particularly heavy borrowers, Type 1 colonies received a much larger portion of their income from loans, Type 2 significantly more, while the dependent colonies appear in much the same stance as Great Britain.

Loans can be productive, but they do generate charges on future income. The political units within the Empire seldom went into default, but that is not true for many of the underdeveloped nations. We have not yet adjusted their budgets for defaults, and those adjustments may alter these conclusions. However, given the raw budget data, we find that the developed countries spent more per capita on dead-weight than Britain, but that the colonies spent less. The underdeveloped parts of the world spent less than the developed but substantially more than the colonies. But while their borrowings were larger, that alone does not account for the difference.

A similar picture emerges from the relative measure, although here there is a definite upward trend over time (but at a decreasing rate), and only self-governing colonies amongst all government categories devoted as large a fraction of their budget to debt service charges as the United Kingdom. When the sample is split in 1900 to take account of changes in British law that permitted certain financial institutions to add colonial bonds to their list of approved securities, borrowing went up but deadweight charges went down. Within the Empire the advantages were never equally shared. The self-governing colonies had always been the heaviest borrowers and their deadweight charges the greatest. While the relative weights do not change after 1900's, the new laws appear to have had the effect of bringing to Type 2 and 3 colonies deadweight charges per dollar of debt more reminiscent of those long enjoyed by the Australasians, the South Africans and the North Americans.

No matter how one feels about Empire as an exploitative venture, it is difficult to believe that membership in it did not give a colony substantially better access to the London capital markets than was enjoyed by countries outside the system. In addition, while it is possible that the British induced their colonies to spend borrowed funds on a myriad of useless projects (a charge that has, for example, been made in connection with the Indian railroads), an examination of the uses of colonial as opposed to noncolonial loans suggests that within the Empire such loans were seldom used to finance current expenditures, but that outside the Empire such was frequently the case

Now we come to the single most important set of expenditures—those dealing with defense. A highly coercive imperial power might wish to relieve its domestic taxpayers of as many of the costs of empire as possible and to shift the burden to its colonies. A comparison across the world indicates that Britain was either unable or unwilling to follow such a policy when it came to defense expenditures—in fact, almost the reverse appears to have been the case

On a worldwide basis, the colonies spent less on defense per capita than Britain, and the independent underdeveloped as well as the developed countries more. In terms of budget percentages, the results are even more marked. Independent countries appear to have spent substantially less than the United Kingdom, but the colonies a very great deal less. Within the Empire the inclusion of every level of government does nothing to change the picture. All are substantially below the United Kingdom both in per capita and percentage terms.

Between colonies there were, however, some differences. While all colonies appear to have benefited from British military protection, some did better than others. India, for example, was the only possession that met the ideal that the British Government had been attempting to establish ever since the French and Indian Wars—that colonies be entirely self-supporting in all areas of domestic government including local defense. Not satisfied with the happy situation on the subcontinent itself, Whitehall used Indian troops in a number of imperial ventures outside India with the expenses continuing to fall on the Indian budget.

The greater a colony's level of autonomy, the more it was able to avoid imperial responsibilities. Thus, the self-governing colonies, recognizing that the British taxpayer would assume the responsibility, spent virtually nothing on defense. Although the British (and to a lesser extent, the Indian) exchequer bore most of the costs of imperial defense, the British Government was able to extract from the dependent colonies contributions towards imperial defense from which the self-governing colonies were largely immune. Thus, the wealthier of their number—Ceylon, Mauritius, Hong Kong and the Straits Settlements—could be made to pay something where, for example, Canada could not.

But Mauritius only had defense costs because it was a "fortress" in an empire it never chose to join and whose shipping it had been elected to protect. Hong Kong harbor, the Colonial Office pointed out, in 1889 was endangered only because it was a coaling station and naval depot from which vessels bound for China and Japan were protected in a "trade which is mainly independent of H. Kong and which is carried on for the benefit of the mother country and the British taxpayer. . . . The 'British taxpayer' who plays so large a part in the Treasury position, is probably the only person who is strongly interested in the defense of Hong Kong. The Island produces nothing, and the defense of it is the defense of British trade. . . ."

While colonial governments paid lip service to the proposition that local defense was a colonial responsibility, regardless of their state of constitutional development, they were often successful in avoiding the cost of actual hostilities. Initial expenses were almost always paid from the treasury chest (a fund of £1,000,000 spread throughout the Empire to cover emergencies), but once the imperial monies had been expended, the British treasury usually enjoyed but small success in recouping its monetary advances.

It is once more apparent that if the purpose of the Empire was to transfer resources from the colonies to the home country, Britain again failed to develop an effective mechanism. Self-governing colonies were able to transfer resources which might have been needed for defense to more desirable goals without placing higher revenue burdens on their citizens. It seems very likely that the rapid accumulation of social overhead capital that marked the histories of late nineteenth-century Canada and Australasia can be traced directly to the defense umbrella provided by Britain. Even India, which contributed so much to imperial defense, paid no more than she might have had she been independent and neither in per capita nor relative terms did the subcontinent's level of expenditures on defense approach that of a typical non-developed independent nation.

All in all one is forced to ask if indeed the Empire was not conceived in a fit of absent-mindedness. Either that or a singular altruistic spirit seems to have pervaded those who occupied the seats of power in Britain, for the system appears to have been ill-designed for exploitation. Such a conclusion appears to be borne out not only by this paper, but also by a series of our parallel studies. One could argue that the budgets give a false picture since the colonies may have been charged artifically high prices for the commodities that they were forced to purchase. However, a study of the Crown Agents (the centralized colonial purchasing agency) suggests instead that the colonies benefited from a certain degree of monopsony purchasing power. To sum up, it appears that there is little evidence of an efficient exploitive mechanism in the colonial governmental sector.

U.K. Savings in the Age of High Imperialism and After

By MICHAEL EDELSTEIN*

This paper summarizes some tentative evidence on the patterns and determinants of aggregate private saving in the United Kingdom over the past century and suggests some implications of these findings for several unresolved issues in the history of *U.K.* accumulation and growth. In particular, three aspects of the literature of modern savings behavior are central to the motivation of this study.

First, although the structure of *U.K.* savings behavior is often crucial to hypotheses concerning the age of high imperialism and capital export, 1870–1913, there has been virtually no econometric testing of savings behavior during this period to help deepen our understanding of these events. For example, J.A. Hobson argued at the turn of the 19th century that *U.K.* business cycle expansions were characterized by an insufficient expansion of consumption demand and too large an offering of savings; according to Hobson the net result was the *U.K.*'s massive capital export of the late 19th century. The basis for this phenomenon was a declining wage share as the cycle expanded and a rising quantity of saving invariant to rates of return. Along with many other 19th century economists Hobson thus believed that the massive *U.K.* capital export was more strongly determined by domestic push forces rather than overseas pull forces. Clearly, any aggregative treatment of these issues necessitates an understanding of the amounts and willingness to save. The shape of *U.K.* savings behavior is also central to another aspect of accumulation in the 1870–1913 period, the long swing alternation of *U.K.* home and overseas investment. Moses Abramovitz and John Williamson argue that this long swing alternation was in large degree the result of long swings in the *U.S.* demand for social overhead capital acting on a "fixed pool" of *U.K.* savings. Brinley Thomas (1954) has suggested the reverse; the long swing motion of the *U.K.* demand for social overhead capital acted on a fixed pool of *U.K.* savings. Neither side of this debate has examined whether the pool of *U.K.* savings was, in fact, fixed.

A second motive for studying the last century of British savings concerns the effects of the changing institutional structure. Keynesian and post-Keynesian models of individual and corporate savings are often linked to the actions of provident households and large corporations with long planning horizons. Virtually simultaneous with the publication of Keynes's *General Theory*, M. M. Postan surveyed the previous two centuries of savings and investment behavior. Although influenced by Keynes's new macroeconomics, Postan argued that the dominance of the provident household and the large self-contained corporation in British savings decisions was a relatively recent development. During the half century before World War I, the saving decision was largely guided by the actions and motives of "pure investors." Unlike the provident and corporate savers, the "pure investor" operated across the full range of British assets at home and abroad, often through public capital markets, with the sole criterion of maximizing expected return. Postan suggested that with the passing of the "pure investor's" importance, the allocation of British saving across the spectrum of assets was less free and, quite possibly, an important cause of Britain's then current economic troubles. Another implication of this shift in the decision locus of the British savings decision, one which Postan skirted, is that the allocation of British income between savings and consumption had become

*Assistant Professor, Queens College, CUNY

less sensitive to expected returns. An important motive for this study is to examine whether the shift in the locus of the savings decision occasioned a decreased sensitivity to expected returns in the determination of aggregate *U.K.* savings, if any such sensitivity ever existed!

A third motive for studying the last century of *U.K.* savings stems from a recent study of *U.S.* aggregate private saving behavior by Paul David and John Scadding. Examining aggregate private saving in the United States since 1896, the authors found a near-constancy in the average gross private saving rate, (*GPSR*) (p. 240). In addition, they presented a strong case for a single, stable behavioral structure determining aggregate private savings during peacetime years. These results may seem to fly in the face of two phenomena many thought central to modern *U.S.* savings behavior: the dramatically increased role of corporate retained earnings in aggregate private saving and the equally dramatic rise of the public sector and its tax base. In a sense David and Scadding are arguing that Postan's 19th century *U.K.* "pure investor" is alive and well in 20th century America, albeit using corporations rather than payout and public placement as the proximal locus of their saving vs. consumption decision.

I

C Feinstein's excellent new study of the *U.K.*'s national income accounts presents a unique opportunity to study these three sets of issues with almost a century of internally consistant data. Because national balance sheets are unavailable before the most recent years and good residual estimates of household and corporate savings only begin in the post-World War II period, the estimator of gross private savings adopted for this study is gross national savings minus government savings. Given that gross national savings is identically equal to gross national investment, Feinstein's investment series can be employed to estimate gross national savings. Figure 1 presents annual values of the ratio of gross private saving to gross national product during peacetime, 1870–1965.[1]

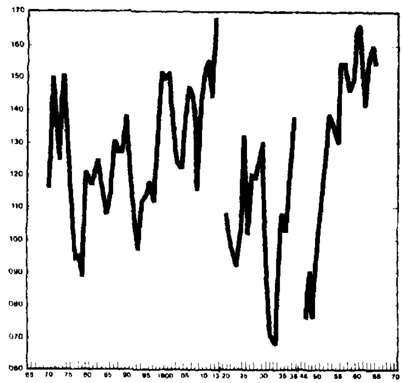


FIGURE 1 U K, PEACETIME, GROSS-PRIVATE SAVINGS RATES, 1870–1965

Setting aside the years 1946–1950 because they were dominated by continuing wartime economic controls, the basic pattern of the peacetime *GPSR* since 1870 involves two rising trends, culminating in peaks just before World War I and the early 1960's.[2] The rising trend from 1870 to 1913 is perhaps best seen as part of the gentle upward drift of the *U.K.* rate of accumulation recorded as far back as 1700 (P. Deane and W. A. Cole). The violent break of the depressed interwar years forms the start of the second rising trend. The data end with the second rising trend hitting average peak values somewhat above the immediate pre-World War II peak rates.

In the pre-World War I segment, off-trend peaks occured in the early 1870's and just before World War I, while the strongest off-trend

[1]The details of constructing the estimator of aggregate private savings may be found in an unpublished manuscript, "Private Saving in the United Kingdom, 1870–1965," available from the author

[2]Estimates of net purchases of consumer durables and their rental value are unavailable for the *U K* except for recent years It is difficult to believe, however, that introducing these two series into our calculations would reverse the general trends noted in the text.

trough occurred in the mid-1890's. Interestingly, the detrended U-shape of the 1870–1913 GPSR is paralleled by the U-shape of the yield on Consols and the realized, price-deflated return on home, long-term negotiable assets (Edelstein).

Although it thus seems fairly certain that the *U.K.* did not experience a unique and stable private savings rate from the 1870's to the present, it is possible that *U.K.* private saving was determined by a unique and stable behavioral structure. Many economists and historians would be skeptical of such a hypothesis. There has been a significant rise in the amount of saving done by large, diversified corporations, presumably with longer decision horizons than either households or the relatively smaller 19th century companies. The role of the *U.K.* government as a producer, accumulator, and insurer has grown enormously, possibly altering the functions of private saving. The *U.K.* income and wealth distributions have become somewhat more equal over the last century, possibly altering the aggregate savings rate by changing the groups who receive the national income. Over the last century the central capital market has become more involved with domestic industry and less involved with overseas portfolio investment. Under the hypothesis that capital tends to move in deep channels, this switch may have had important effects on aggregate behavior. On trend, *U.K.* birth rates have tended to fall, and the percentage of the population in the prime working years to rise. Lifecycle savings considerations would argue that the saving rate might have been altered by these changes.

Of course, it is possible that these institutional, distributional, and demographic changes only affected the proximal locus of the savings decision, or perhaps how savings were funneled to investment projects, not the ultimate determinants of private saving behavior. Hypotheses which stress psychic or socioeconomic class determinants of savings behavior might argue that psychic or class factors change very slowly, themselves being products of very slowly evolving cultural and historical forces (e.g., David and Scadding). We must ask, however, *whether the mid-20th century U.K. capitalist is as psychically secure as his or her mid-19th century predecessor. With regard to the balance of class forces, has not the last century seen somewhat greater show of political power and economic returns flow into the hands of U.K. labor?*

The approach employed here to investigate these hypotheses begins with fitting a wide variety of modern savings models to *U.K.* data for the peacetime years since 1870. Estimates are made of the parameters of the Keynes, Duesenberry-Modigliani, Brown-David-Friedman, Mincer, Kaldor, and Ando-Modigliani models of savings behavior in their original forms and numerous variants.[3] Each model is estimated by dividing both sides of their standard level's formulation by income *(Y)* to avoid problems of heteroscedasticity; the dependent variable is thus the *GPSR*. Based on the standard error of the regression and Durbin-Watson statistics, the best fitting models are the Brown-David-Friedman, Mincer, and Kaldor Lagged (Table 1, Panel A).[4]

Next, the yield on Consols is introduced to test the sensitivity of aggregate private savings to expected returns.[5] With the one exception of an unlagged version of the Kaldor model, the coefficient on the interest rate term is insignificant in all simple formulations of the modern savings models noted above. Setting aside for the moment the possibility that structural shifts are affecting these results, it is possible that the interest rate effect is hidden by the failure to

[3]In the functions incorporating Koyck-type lag structures, $((Y - S)/Y)_{-1}$ is employed to embody the idea that the short-run marginal propensity to save *(mps)* is larger than the long-run *mps*. A direct proof of this assumption may be found in the hybrid Friedman-Ando-Modigliani model introduced later.

[4]The standard error of the regressions of the rejected models were all greater than .0195. Their Durbin-Watson statistics all indicated statistically significant negative serial correlation.

[5]The interest rate employed here is not deflated by price expectation, $\hat{p}$. At an early stage an estimator of $\hat{p}$ based on long lags was introduced and the interest rate coefficient was insignificantly affected. Warren Weber (1975) found that when r and $\hat{p}$ are separately introduced into a consumption function, the latter's effect is insignificant. This finding parallels our own that the r coefficient is stable across the 1873–95 deflation and 1896–1913 inflation.

TABLE 1—ESTIMATES OF PARAMETERS OF AGGREGATE SAVINGS BEHAVIOR, 1872–1913, 1922–38, 1951–65

		S.E.R.	D.W.
	A. Some Simple Models[a]		
(1) Brown-Davis-Friedman[b]	$-23.8\ (1./Y) + .77 - .72\ ((Y\text{-}S)/Y)_{-1}$	.0141	1.90
	(2.5) (12.) (9.3)		
(2) Mincer	$-26.8\ (1./Y) + .76 - .24\ LFE - .42\ ((Y\text{-}S)/Y)_{-1}$	.0118	1.79
	(3.8) (13.) (5.6) (5.1)		
(3) Kaldor Lagged	$-78.4\ (1./Y) + .66\ W/Y + .79\ NW/Y - .64\ ((Y\text{-}S)/Y)_{-1}$	.0124	1.74
	(4.0) (9.4) (12.) (8.3)		
	B. Some Hybrid Models		
(4) Classical	$.56 - 169\ 2\ (W/NW)/Y + 19.6\ r/Y - .42\ ((Y\text{-}S)/Y)_{-1}$	.0118	1.56
	(8.1) (5.3) (3.0) (4.9)		
(5) Friedman-Ando-Modigliani	$-42.7\ (1./Y) + .21\ Y^P/Y + .64\ Y^t/Y + 10.0\ r/Y - .0237\ (A_{-1}/Y)$	.0119	.64
	(3.1) (20.0) (8 6) (1.9) (9.8)		
	C. Evidence of Structural Shift		
(6) Brown-Davis-Friedman	$-36.0\ (1./Y) + .63 - .55\ ((Y\text{-}S)/Y)_{-1} - .016\ ((Y\text{-}S)/Y)_{-1}D3$	.0133	1.81
	(3 7) (8.2) (6 1) (3 1)		
(7) Kaldor Lagged	$-85.1\ (1\ /Y) - 398.0\ (1\ /Y)D3 +\ 49\ W/Y -\ 18\ (W/Y)D3 + .81\ NW/Y$		
	(5 2) (5 1) (6.2) (4.3) (12)		
	$-\ 56\ ((Y\text{-}S)/Y)_{-1} +\ 76\ ((Y\text{-}S)/Y)_{-1}D3 + .05\ ((Y\text{-}S)/Y)_{-1}D4$	.0100	1.67
	(7.8) (5 0) (3 1)		
(8) Classical	$40 - 194.0\ (W/NW)/Y - 24\ 3\ ((W/NW)/Y)D3 + 23\ 8\ r/Y - .23\ ((Y\text{-}S)/Y)_{-1}$	.0106	1.48
	(5.6) (6.6) (4 2) (4 0) (2.6)		
(9) Friedman-Ando-Modigliani	$-245\ 9\ (1./Y) - 194\ 2\ (1\ /Y)D3 + 406.6\ (1\ /Y)D4 + .35\ Y^P/Y -\ 17\ (Y^P/Y)D4$		
	(12) (3 7) (5 6) (16) (8 2)		
	$+\ 76\ Y^t/Y + 54.6\ r/Y -\ 0475\ (A_{-1}/Y) +\ 0142\ (A_{-1}/Y)D3$	.0064	1.56
	(18) (10) (8.3) (3.1)		

[a]Abbreviations: Y = gross national product, $Y^P = (Y + Y_{-1} + Y_{-2})/3.0$; $Y^t = (Y^P - Y)$, W = income from employment; $NW = (Y - W)$, A = (real net reproducible capital stock) + (net accumulation of overseas assets, deflated) + (funded and unfunded debt, deflated); LFE = labor force/employment, r = yield on consols; $D3$ = dummy variable for the years, 1922–38; $D4$ = dummy variable for the years, 1951–65.

[b]Note that all equations are estimated in ratio form, i e , the dependent variable is S/Y (= $GPSR$). The term $(1./Y)$ is thus necessary to capture the possibility of a constant term in the levels form The bracketed figures under the regression coefficients are t-statistics.

comprehend some other phenomena along with the interest rate. The exception of the unlagged Kaldor model (not reported here) suggests that these other phenomena might be connected with the variation of the factor income distribution and, in turn, this suggests that the explicit variation of permanent and transitory income estimators also might bring out the interest rate effect. Two hybrid equations are thus introduced: a "Classical" model in which aggregate private savings are a function of current and lagged values of the level of income, the ratio of wage to nonwage income, and the interest rate, and a "Friedman-Ando-Modigliani" model in which aggregate private savings are a function of permanent and transitory income, private wealth and the interest rate (Table 1, Panel B).[6]

These hybrid equations achieve much better

[6]The Friedman-Ando-Modigliani model may be viewed as crude in two ways: the simplistic idea of how permanent income expectations are formed and the assumption that only lagged values of permanent income effect current savings. The introduction of a Koyck-type lag structure involves a nonlinearity (El-Modadem). Nonlinear estimators were prepared and the equation for the full period exhibited an excellent fit. Unfortunately, computational difficulties prevented estimation in the early subperiods.

fits over the century of data than the "purer" models and, importantly, the interest rate coefficients are positive and statistically significant.[7] The calculated long-run elasticities of the interest rate with respect to gross private savings are .37 in the Classical model and .11 in the Friedman-Ando-Modigliani model. These elasticities are perhaps somewhat low but if one assumes that there is simultaneous equations bias interfering with our ordinary least squares estimates of the interest rate coefficient, the strong inference is that these elasticity estimates are underestimates of their true values! Investment demand is the most likely source of bias and, typically, the interest rate term in that relation has a negative coefficient.

The negative sign on the *W/NW* ratio is quite interesting. This ratio is countercyclical on average; the share accruing to wage incomes tends to fall as the business cycle expands and rise as the business cycle contracts. This means that the *U.K. GPSR* tended to rise and fall with the business cycle Clearly, this model could be reasonably interpreted to embody of the underconsumptionist ideas on the operation of the 19th century business cycle (e.g., Sismondi, Marx, Hobson). It could be that *W/NW* fell in a stabilizing fashion, generating the extra profits and savings in the upswing and thus mitigating inflationary pressures. But, there is no theory which argues that the reduced consumption and augmented savings need be stabilizing. In the Friedman-Ando-Modigliani model, the explanation of the *GPSR* moving independently of the rate of return is the increased component of positive transitory income which appears as the cycle expands.

As yet untested is the central question of whether the last century of *U.K.* private savings behavior was determined by a unique and stable behavioral structure. A Chow test for the equality of coefficients across the 1872–95, 1896–1913, 1922–38, and 1951–65 subperiods was performed for each model. Based on these tests it would be difficult to make a strong case of a unique and stable behavioral structure spanning the peace time years of the last century.[8] The Brown-Davis-Friedman and Classical models come close to evincing homogeneity across the various subperiods, but they do not pass at the .05 significance level. The Kaldor Lagged, Mincer, and Friedman-Ando-Modigliani models are somewhat less stable than this first group and the unreported models were even less stable. Nevertheless, some models manifested stability within the 1870–1913 and the 1922 . . . 1965 periods. Rather than speculate about this result based on the equation-wide Chow test, it seems worthwhile to proceed to an examination of exactly which parameters shifted.

II

Towards this end, each parameter of the Brown-Davis-Friedman, Kaldor Lagged, Classical, and Friedman-Ando-Modigliani models was tested for structural shift between the 1872–95, 1896–1913, 1922–38, and 1951–65 subperiods. This was achieved by multiplying each variable by dummies for the latter three subperiods, thus, these new variables were constructed as contrasts. The statistically insignificant contrasts were eliminated, leaving equations 6–9 of Table 1, Panel C.

As noted earlier, Postan's argument that the 19th century saver was a "pure investor" and that this type of saver was succeeded by provident householders and corporations with very long planning horizons around the time of World War I might be construed to imply the interest rate elasticity declined at that time. Interestingly, if one assumes that the *only* structural shift in aggregate *U.K.* private savings behavior since 1870 was a change in the interest rate elasticity, this is exactly the result confirmed by econometric testing in evidence presented elsewhere. From a calculated elasticity of greater than unity in the pre-World War I years, the interest rate elasticity was cut in half during the interwar period, and disappeared in the post-World War II years. Ironically, this result confirms the classical political economists during their century, Keynes and Postan

[7]Note that the Classical model has no (1 /Y) term The coefficient is insignificant for the full century and all subperiods and is thus dropped, the implication is that the marginal and average propensities to save were the same This is the only model yielding this strong result

[8]The Chow test F-statistics for the first five equations of Table 1 are (1) B-D-F, 2 29 (9,64); (2) Mincer, 3 88 (12,60); (3) K-L, 6 04 (12,60), (4) Classical, 3 83 (15,54), (5) F-A-M, 7 04 (15,54)

for the interwar years, and the post-Keynesian econometricians in their post-World War II modeling. The problem with this story is that it ignores other forms of structural shift. Indeed, when other loci of structural shift are admitted to the econometric analysis as in the hybrid equations 8 and 9 of Table 1, the interest rate effect appears to have remained strong and positive throughout the last hundred years with a calculated elasticity of around .6–.7![9]

If the coefficient on the interest rate has remained fairly stable over the last hundred years, which parameters did shift and to what effect? The years from 1872 to 1913 saw no statistically significant shift in any parameter. However, during the interwar period, there are indications that a number of income and wealth parameters shifted. Importantly, these changes seem explicable in terms of the depressed conditions and prospects of the period rather than any institutional shifts. The Classical and Friedman-Ando-Modigliani models give differing explanations, however.

Although equations (6) and (9) present excellent evidence that the short-run *mps* out of total income remained stable across the last century, the models incorporating factor income phenomena, the Kaldor Lagged and Classical, suggest a downward shift in this parameter in the interwar years. In the Kaldor Lagged model (7) the coefficient on $(W/Y)D3$ directly indicates a fall in the short-run *mps* out of wage income. The short run *mps* out of nonwage income apparently held constant. In the Classical model (8) the coefficient on $(W/NW)/Y$ becomes more negative, implying that in the interwar upswings (downswings) relatively less saving (more dissaving) occurred than other subperiods. The size of the coefficient on $((W/NW)/Y)D3$ is small, however suggesting this factor is less important than implied by the Kaldor Lagged model. Furthermore, the (W/NW) variable may be a proxy for investment, as well as savings behavior. It is possible that not only were investment demand profit expectations inordinately low during the interwar decades but businesses were less responsive to those that did exist; W/NW may proxy these phenomena. Thus, within the Kaldor/Classical world, the fall in the *mps* out of wage income gives some explanation of the fall in the interwar *GPSR* but it appears to be a minor influence relative to the impact of lowered investment demand.

Paradoxically, in the Friedman-Ando-Modigliani model (9) the only interwar structural shift apart from the shifting constant term was a change in the wealth coefficient tending to raise saving, all else held constant. Generally, wealth manifested a negative relationship with savings during the last century, presumably in part reflecting the life-cycle consumption purposes of such accumulations. The depressed interwar decades seem to have weakened this effect so that those increases in private wealth which did occur resulted in a relatively smaller increase in consumption and more savings than other subperiods, all else held constant. Since wealth is far more concentrated than incomes, one might infer that it was changes in the long-run expectations of the well-to-do which altered the parameters of savings in the interwar years, contrary to the Kaldor-Classical results which imply that it was wage earners who were inordinately affected. Because of difficulties in estimating more complete models involving investment demand equations, a resolution of this difference and a more precise explanation of the interwar fall in the *GSPR* is impossible at this stage.

In the post-World War II period many parameters returned to their pre-World War I levels. The Friedman-Ando-Modigliani equation (10) suggests that the *mps* out of permanent income fell. Why this should be the case is not easily resolvable. One plausible hypothesis is that with post-World War II governments socializing so many "rainy day" and retirement motives for savings, the need for saving out of permanent income was reduced.

III

Very briefly then, what are the major implications of these empirical findings for the several issues raised in the introductory passages? First, because savings behavior appears to have been sensitive to interest rates, the long-swing alternation of *U.K.* home and overseas investment, 1870–1913, seems far less dependent

[9]Contra Weber (1970, 1975) whose findings implied a negative elasticity for the United States in the mid-20th century

upon the notion of a fixed pool of savings than previously suggested. Indeed, at both the very beginning and end of the 1870–1913 period when the *GPSR* hit levels well above trend and interest rates were also at their highest for the 1870–1913 period, both home and overseas investment were being accommodated and, at least to a degree, not in need of alternation. Second, it seems clear that shifts in the factor income distribution affected savings over the cycle. The shift towards a lowered wage share as the cycle expanded tended to raise the *GPSR* (lower the consumption-income ratio). It is possible that some form of underconsumption could have resulted near the cycle peak. It thus appears that there is plausibility to both J.S. Mill's conjectures on the interest rate sensitivity of *U.K.* savings and Hobson's emphasis on the role of income distribution factors determining *U.K.* aggregate private savings. For example, in the 1904–13 expansion of foreign investment from 2.63 percent of *GNP* to 9.24 percent, domestic investment fell from 9.71 percent to 6.29 percent and the *GPSR* rose from 12.19 percent to 16.76 percent. Clearly there is some substitution but it is not pound for pound as implied by a fixed pool. Employing the Classical model, the expansion in the *GPSR* is roughly attributable to both a lower *W/NW* and a higher *r*.

Third, since the interest rate coefficient appears to have remained stable throughout the last century, one implication might be that at least in some institutional garb, Postan's 19th century "pure investor" is alive and well. Fourth, the structural shift in the parameters determining aggregate private saving in the interwar period seem much more related to the depressed expectations and conditions of the two decades than they do to institutional shifts of the type he hypothesized. Indeed, the finding of no structural shift, 1870–1913, buttresses this conclusion because the 1896–1913 period occasions the first major industrial merger movement. Fifth and finally, David and Scadding's ultrarational savers who subsume private and government decisions to maintain a steady *GPSR* and a unique, stable structure of savings behavior are not in evidence in the *U.K.* However, to the extent that the models of *U.K.* savings behavior which aggregate over household and corporate savings decisions have been here found strong and stable over plausible, historical epochs, one might hypothesize that savers in the *U.K.*, as well as the United States, attempted to subsume household and corporate activity in their decisions.

REFERENCES

Moses Abramovitz, "The Passing of the Kuznets Cycle," *Economica*, Nov. 1968, N. S., *35*, 349–67.

Paul David and John Scadding, "Private Savings: Ultra Rationality, Aggregation and 'Denison's Law,'" J. Polit. Econ., Mar. 1974, *82*, 225–50.

P. Deane and W. A. Cole, *British Economic Growth, 1688–1959*, Cambridge 1969.

Michael Edelstein, "Realized Rates of Return on U.K. Home and Overseas Portfolio Investment in the Age of High Imperialism," *Expl. Econ. Hist.* Summer 1976, *13*, 283–329.

A. M. El-Mokadem, *Econometric Models of Personal Saving: U.K., 1948–1960*. London 1973.

C. Feinstein, *National Income Expenditure and Output of the United Kingdom, 1855–1965*, Cambridge 1972.

J. A. Hobson, *Imperialism: A Study*, London 1902.

J. M. Keynes, *The General Theory of Employment, Interest, and Money*, New York 1936.

M. M. Postan, "Recent Trends in the Accumulation of Capital," *Econ. Hist. Rev.*, Oct. 1935, *6*, 1–12.

Brinley Thomas, *Migration and Economic Growth*, Cambridge 1954.

Warren E. Weber, "The Effect of Interest Rates on Aggregate Consumption," *Amer. Econ. Rev.*, Sept. 1970, *60*, 591–600.

———, "Interest Rates, Inflation, and Consumer Expenditures," *Amer. Econ. Rev.*, Dec. 1975, *65*, 843–58.

Jeffrey Williamson, *American Growth and the Balance of Payments, 1820–1913*, Chapel Hill 1964.

IMPACT OF RECENT DEVELOPMENTS IN PUBLIC FINANCE THEORY ON PUBLIC POLICY DECISIONS

Some Lessons from the New Public Finance

By JOSEPH E. STIGLITZ AND MICHAEL J. BOSKIN*

In the last few years, there has developed a large literature, sometimes referred to as the "new public finance," providing a quantitative analysis of a number of traditional problems within the field. This paper is concerned with surveying, or interpreting, what can be learned from this literature; and our belief is that it has taught us a great deal. We concern ourselves here not so much with the derivation of precise formulae, e.g., for optimal tax rates, but with the more general lessons which have emerged. Some of these are of a philosophical sort—how we ought to think about designing tax structures; some are of a negative sort—pointing out fallacies in traditional arguments or the lack of generality of previous results; finally, some are of a positive sort—deriving the conditions under which a particular tax structure or provision would be desirable. We shall present an example of each type of lesson in turn.

I. The Philosophic Basis of the New Public Economics

Two strands may be identified in the recent literature. The first is explicitly normative—it takes some criterion, usually the utilitarian objective of maximizing the sum of utilities; makes some assumptions about the structure of the economy, including the set of instruments available to the government; and then derives from them some propositions concerning the optimal tax structure.[1] It explicitly recognizes the second (or third) best nature of the problems being discussed; indeed, this literature probably represents the most significant body of work in "second best economics." This is important, because, as we shall show later, much common reasoning on tax problems is based on a misapplication of first best economics.

The problem of taxation in an economy such as ours is viewed as a problem of indirect control of imperfectly observable variables (see, for instance, A. Atkinson and Stiglitz, 1976); the government, for instance, might like to exempt "necessary" medical expenses, but finds it difficult (costly) to distinguish between these and "unnecessary" medical expenses; we might like to have an ability tax (which presumably would be nondistortionary), but we can only observe income, a compound of ability and effort; we might like to distinguish between wage and capital income in the unincorporated sector, but there is no obvious way of doing so; we might like to tax the output "automobile services" but it is much less expensive to monitor the inputs (gasoline and new cars).

This way of looking at tax problems is important because much of what otherwise might appear to be capricious, distortionary, or inequitable may at least make sense, and perhaps even be judged to be desirable.

The utilitarian framework has the advantage

*Stanford University and Oxford University, and Stanford University and National Bureau of Economic Research, respectively. This research was supported by the U.S. Treasury Department.

[1]See, for example, William Baumol and David Bradford, 1970; Peter Diamond and James Mirrlees, 1971; Arnold Harberger, 1964 and A. Atkinson and Stiglitz, 1972.

of providing a simple, unified, reasonably flexible ethical basis for judging among tax systems; for instance, by positing social welfare functions with different degrees of elasticity of substitution among the utilities of individuals, one can consider, at the one extreme, the Rawlsian criterion of maximizing the utility of the worst off individual, and at the other, the Benthamite criterion of adding up utilities. The traditional approach has involved "listing" criteria, e.g., horizontal equity, vertical equity, administrative costs, *without providing any criterion for trading off among these objectives.*

In fact, the principle of horizontal equity, at least as it has usually been formulated as equal tax treatment of equals (*ex post* equality, as opposed to *ex ante* equality, where they all have the same chances), is inconsistent with utilitarianism; i.e., in a wide variety of cases, maximizing the sum of (expected) utilities necessitates random taxation (Stiglitz 1976a); as a practical matter, this might provide a justification for the random enforcement of taxes. It perhaps should be emphasized that this result is not an oddity; the conditions under which random taxation would be desirable are reasonably weak (e.g., with separable utility functions, all that is required is greater than unity risk aversion). The traditional "concavity argument," e.g., of Abba Lerner and Paul Samuelson (i.e., that because of diminishing marginal utility, social welfare is increased by equating incomes of people who are otherwise the same) fails in the context of the second best problems on which the new economics focuses.

The second strand in the literature is more descriptive in character. It shares with the first strand its concern for the general equilibrium analysis prevalent in the earlier literature and its emphasis on "distortionary" analysis—its belief, or in the case of the empirical work, its verification that many of the relevant elasticities in the economy are far from zero (see Boskin, 1977, James Heckman 1974 and Martin Feldstein 1975). This is important: not only are the quantitative effects misjudged by assuming zero elasticities or ignoring general equilibrium effects, but the qualitative aspects of the desirable tax structure may be altered. For instance, an inheritance tax may well—when account is taken of the effects of the tax on saving, the effect of saving on the long-run capital stock, and the effect of the capital stock on the distribution of incomes—increase the degree of inequality in the wealth distribution rather than decrease it (Stiglitz 1966); even apart from the capital accumulation effect it may increase the degree of inequality in consumption (Stiglitz 1976); the social security system may similarly have adverse effects (Feldstein 1974); the elimination of the provisions for charitable deduction may well decrease total expenditures on public type goods (Boskin 1976a and Boskin and Feldstein 1977).

The concern for the general equilibrium effects of a tax policy has led to the introduction or reemphasis of at least two concepts in assessing alternative programs. Stiglitz, arguing that, at least in many cases, the capital accumulation effects of a tax can be offset by government monetary policy, has suggested the use of "balanced growth path incidence analysis" (1976b); in analogy to Richard Musgrave's use of balanced budget analysis, the capital labor ratio, rather than national income or the budget, is held constant. Feldstein has, however, demonstrated that the size of some programs, in particular social security, is so large that its effects cannot be easily offset by monetary policy.

The other concept, the use of which is now written into law, is that of tax expenditures: the loss of revenue due to a particular provision. We have three major objections to this concept. First, as presently formulated, the measurement of foregone revenue implicitly assumes zero elasticities; the estimates of aggregate tax expenditures are correct only when one contemplates eliminating *all* deviations from taxing real economic income *simultaneously and* if the factors of production are in *perfectly inelastic supply* (which Boskin 1977 and Heckman 1974, among others, demonstrate is not the case). Further, the estimates for particular so-called tax

preferences are often extremely inaccurate. For example, if the tax law allows a deduction for charitable contributions, it is not correct to argue that abolishing the deduction will increase tax revenue by (the summing over all contributors who itemize deductions) the product of the marginal tax rate and the amount currently given to charity. The amount of resources flowing into each such "tax expenditure" category reflects the tax treatment of that category as well as others. Since the charitable deduction reduces the price for a dollar of charitable contributions from \$1 to $\$(1-t)$, where t is the marginal tax rate, any price elasticity at all in charitable giving would imply that abolishing the deduction would also reduce charitable contributions. Take the case of a family with a marginal tax rate of 20% which currently gives \$300 a year to charity. The tax expenditure budget counts .2 times \$300, or \$60, as a tax expenditure. Yet abolition of the deduction implies a 25 percent price increase; with the elasticity of -1.2 estimated by Feldstein (1976), contributions fall to \$210, and at the other extreme the "revenue foregone" is only \$42 if the extra \$90 does not flow into taxable income. The tax expenditure budget thus overestimates the revenue loss by more than 40 percent! While it may not be inaccurate to argue that abolition of some preferences would increase taxable income by the amount now assumed in the tax expenditure budget, in many cases it is likely to be *very* inaccurate. For example, it would be heroic to assume that "full" taxation of capital gains would increase taxable income by anywhere near the tax expenditure budget's estimates.

Second, as pointed out recently by Feldstein (1975b) government spending on an activity such as charity may decrease private spending on the commodity. The rationale for the tax expenditure budget is that the government could collect the foregone revenue and spend it on the "preferenced" commodity directly; it is obvious, however, that if each government dollar crowded out a private dollar, a nonzero pure substitution elasticity implies that the government would have to spend *more* than their estimated revenue loss to provide the equivalent total expenditure on the commodity. Third, the tax expenditure concept suffers from a further defect: the legislation implicitly assumed that the "natural" tax base is income, broadly defined; as we shall argue below, there is little justification for this. That is, to know what is being "exempted" from taxation one needs to know what "ought" to be taxed. By using income defined in a broad way, the legislation focuses discussion on particular aspects of the tax code. (For instance, one might also argue that the failure to allow depreciation on human capital is a negative tax expenditure. See Boskin, 1976b.)

The recent quantitative literature has also emphasized the importance of looking at the detailed structure of the tax code. For instance, because of the interest income exemption, the relevant marginal cost of capital for a corporation may be significantly different from the average cost; indeed, in the absence of uncertainty and with the appropriate depreciation provision,[2] the corporation tax would be nondistortionary (equivalent to a pure profits tax); in some cases, the marginal cost of capital might even be less than the before tax rate of interest. (See Stiglitz, 1973). With true economic depreciation, a corporation tax without interest income exemption would also be nondistortionary (see Samuelson, 1964; for a summary, see Stiglitz, 1975). These aspects of the tax structure only become clear upon a detailed quantitative analysis.

II. Choice of the Tax Base

One of the most controversial issues in the literature on public finance is the choice of the tax base. The issue takes on a number of forms:

a) The breadth of the tax base: Should income be broadly defined, and special treatment of medical expenses, charity, etc. be eliminated?

b) Consumption versus income: at least since Fisher, there has been widespread senti-

[2]In this case, equivalent to immediate write off of the investment

ment among academic economists that consumption provided a better base than income.

c) Negative incomes tax versus specific subsidies (e.g., for food, housing, etc.). One of the contributions of the recent literature (see Atkinson and Stiglitz, 1976) is its emphasis that subsidies and taxes are really symmetrical, and therefore the question of the correct "subsidy base" is really the same as the question of the correct "tax base."

Much of the conventional wisdom on this—that the consumption tax is desirable because it does not interfere with the intertemporal allocation of income (the marginal rate of substitution between consumption today and tomorrow remains the same as the marginal rate of transformation) and the subsidies on food and housing are inefficient in distorting the allocation of expenditure among commodities—is only correct under certain conditions; the argument that fewer distortions are better than more distortions is simply wrong. Faced with a second-best problem, the use of first-best welfare economics may lead to seriously erroneous conclusions; a detailed analysis of the problem, taking account of the other distortions in the economy, is often necessary.

For example, Atkinson and Stiglitz have shown that if the utility function is separable between labor (leisure) and goods, if the source of inequality is in ability, and if the consumption tax is chosen optimally, then the consumption tax is the only tax to be imposed; there should be no commodity taxation and no taxation of interest income. If an optimal linear consumption tax is imposed, then this result holds only approximately, i.e., there should be "small" differential taxes or different commodities, which depend on third and higher derivatives of the utility functions.

However, when utility is not separable, a consumption tax is not the only desirable tax; but whether there should be a subsidy or a tax on interest income is a moot question; if we simplify the analysis by assuming a two-period life cycle model, with individuals working and consuming the first period, and only consuming the second, with intertemporal separability in the utility function, then there should be an interest income subsidy (tax) if consumption and leisure are Edgeworth substitutes (complements).

These recent results thus not only cast doubt both on the generality of the conventional wisdom in favor of consumption taxes, and the significance of the earlier literature on optimal indirect taxation, in which only commodity taxation at constant rates (as opposed to progressive consumption or income taxes) was allowed, but point out the relevant empirical information (e.g., whether utility is separable between leisure and goods) necessary to establish the desirable tax structure.

Much recent empirical research has established a non-negligible interest elasticity of saving and wage elasticity of labor supply (see Boskin, 1977 and Hurd, 1976) Feldstein 1975b has analyzed the desirability of consumption taxation in light of this evidence and concludes that a decrease in capital income taxation is desirable

III. Special Provisions: Medical Allowances

The provisions for deductibility of certain expenses and tax credits for others may be analyzed within the same framework. A tax credit is equivalent to a proportional subsidy for the given good (at least for all individuals paying sufficient taxes), while deductibility lowers the effective cost to individuals at a higher marginal bracket more than those at a lower marginal bracket.

The conventional wisdom on this is that deductability provisions are undesirable (relative to tax credits) because the differential prices paid by individuals result in a distortion, and that because the price is lowered more for the rich than for the poor, deductibility provisions are inequitable Neither argument is convincing: the first represents another misapplication of first-best economics to a second-best problem. The second assumes what is to be proven: What is the appropriate "equitable" tax base? One could equally well argue that the appropriate tax base is real income, and that nominal income minus medical expenses is a better proxy for real in-

come than nominal income alone. In that case, there is a presumption for tax deductibility, rather than for a tax credit.

In the analysis referred to in Section II, individuals differed only with respect to their earning capacity. If that were the case, then neither tax deductibility nor a tax credit would be desirable (under the assumption of separability). The argument then for either must be based on a recognition of individual differences with respect to medical needs, and a belief that an ideal tax system (a perfect screening system) would differentiate the tax burden on people with different medical situations. The argument may be put in either ability to pay or utilitarian terms: that necessary medical expenses are a subtraction from the individual's ability to contribute to the support of government services or that they raise the marginal utility of income (since they represent a subtraction from "enjoyable consumption") An ideal tax system would thus relate taxes to ability, A, and health needs:

$$T^* = T(A,H)$$

where H is health, with $T_H < 0$. But neither A nor H is easily observable We shall analyze the desirability of the alternative programs within the utilitarian framework.

Assume that the demand for medical services of any given individual is perfectly inelastic; a particular amount is required just to survive; any amount beyond that has zero utility. Differences in these required medical expenses are the only way in which individuals differ with respect to H. Then medical expenses would be a perfect surrogate for H. If we assume that the individual's indirect utility function can be written in the form

$$V(p,I-H)$$

then the appropriate tax treatment would be to give a 100 percent tax credit, assuming H, I, and A are uncorrelated, and assuming that the government can impose a uniform lump sum tax (subsidy).

In the realistic case, the demand for health is not completely inelastic (see Feldstein, 1974a), so that actual expenditures are only a surrogate for medical needs. But it has an even more important implication: allowing a full credit would obviously be extremely distortionary, leading all individuals to demand medical services up to the point of satiation. Thus, any tax system must allow for only partial payment by the government of medical expenses.

Indeed, we can view the tax deductibility provision as a form of partial insurance for medical expenses; the partial nature of the insurance arises from the same kinds of considerations which lead to co-insurance in conventional insurance policies—moral hazard. Individuals in different income brackets are, however, offered different insurance policies, and an insurance policy with a nonconstant co-insurance provision. To the extent that they face the same risks (in dollar equivalent terms), have the same demand elasticities, and have the same relative risk aversion, since the risks are smaller relative to income (wealth) for the wealthier, the return to having the insurance is smaller while the deadweight loss is the same, leading to some presumption for the wealthier to have less insurance, i.e., the government should allow them a lower tax credit rate (a lower percentage of their medical expenses being deductible). Moreover, if there is diminishing marginal utility of consumption of nonmedical goods, then it would seem desirable to have a larger fraction of large medical expenses insured than of small medical expenses. The tax deductibility provision works in just the opposite way, providing smaller marginal co-insurance rates (for a given income) as expenses go up. This provides some further argument against deductibility over a tax credit.

But income may respond to health needs as well; to compensate for greater medical needs, an individual may be induced to work harder, in which case higher incomes do not represent higher levels of "enjoyment," but only greater needs. In that case, deductibility is preferable to

a credit, since income net of medical expenses is clearly a better measure of welfare than just income alone.[3]

This is illustrated by the following simple model. Let the utility function be

$$(1) \qquad W = U(C) + Z(M,H) - VL$$

where C is "effective" consumption, M is medical expenditures, H is health, and L is labor supplied. Z represents the "direct" utility of medical expenditures.

In general, effective consumption will be a function of health, income, and expenditures on medical care. One simple specification is[4]

$$(2) \qquad C = (I - M) + (M - H)$$

C equals expenditure on nonmedical items, plus all expenditures on medicine in excess of health needs, H. For simplicity, assume I consists of only labor income minus tax payments,

$$(3) \qquad I = WL - T(WL - \lambda_1 M) + \lambda_2 M.$$

T is the tax on income after medical deductions, $T' > 0$, $T'' > 0$ (if the tax is progressive), λ_1 is the percentage deductability allowance and λ_2 is the percentage tax credit. Thus the individuals first-order conditions are

$$(4a) \qquad U'W(1 - T') = V$$

$$(4b) \qquad U'(\lambda_1 T' + \lambda_2) + Z_M = 0.$$

The government wishes to choose λ_1 and λ_2, the percentage deduction and credit allowances, to maximize

$$(5) \qquad \int W(W,H)dF(W,H)$$

[where F is the distribution function of individuals by wage rates (ability) and health needs (H)] subject to

$$(6) \qquad R = \int [T(I - \lambda_1 M) - \lambda_2 M]df.$$

We thus obtain, letting μ be the Lagrangian multiplier associated with (6),

$$(7) \qquad \int M\left[U' - \mu\left(1 - \frac{S}{1-S}\epsilon\right)dF \lesseqgtr 0 \quad \text{as} \begin{cases} \lambda_2 = 0 \\ 0 \le \lambda_2 \le 1 \\ \lambda_2 = 1 \end{cases}$$

$$(8) \qquad \int T'M\left[U' - \mu\left(1 - \frac{S}{1-S}\epsilon\right)dF \lesseqgtr 0 \quad \text{as} \begin{cases} \lambda_1 = 1 \\ 0 \le \lambda_1 \le 1 \\ \lambda_1 = 0 \end{cases}$$

where ϵ is the price elasticity of the demand for medicine and where S is the marginal subsidy rate Straightforward calculations establish that (provided $Z_{MH} > 0$) if $\lambda_1 = 0$

$$\frac{\partial C}{\partial H} < 0, \quad \frac{\partial M}{\partial H} > 0, \quad \frac{dI}{dH} > 0.$$

Increasing health needs increase medical expenditures and lower effective consumption

If the only source of variability in M arose from variations in health needs, H, then it is easy to establish (assuming the price elasticity of medicine does not increase with health needs) either only tax deductibility should be allowed or both a credit and deductibility should be employed, but a tax credit should never be used alone[5]

If, on the other hand, labor is inelastically supplied (I is fixed), equations (7) and (8) remain unaffected; but now, if H is the only source of

[3]We are concerned here more with the considerations entering the choice between a credit and a deduction than with the realism of the model

[4]Other specifications, e.g., $C = I - M$, have similar results

[5]Assume $\lambda_1 = 0$ Then when (8) is nonpositive, (7) is negative

variability, then a deductibility provision should never be used (except if a full credit ($\lambda_2 = 1$) is used in addition, which, since this is equivalent to giving M away, it never will be).[6]

Thus, whether the insurance considerations or equity considerations dominate depends on the responsiveness of income to health considerations.

Even this analysis, we suspect, overstates the case for the desirability of a tax credit. Assume that different individuals had inelastic demands for medical care, but that their demands were only partially correlated with true health needs. Assume M and H are jointly normally distributed, with correlation coefficient ρ, which is independent of I. Assume a social welfare function of the form

$$W = \int [U(I - H - T(M,I))dF$$

where F is the distribution function of individuals over H, I, and M, and where $T(M,I)$ is the tax function which we assume to be of the form $T = T(I - \lambda_1 M) - \lambda_2 M$. In particular, we assume U is quadratic. Then it is easy to show maximizing social welfare requires a *deduction* from income of an amount ζM.

[6]When (8) equals zero, (7) is positive, this follows upon observing that

$$\frac{dU'}{dH} = -U''\left(1 - T'\lambda_1 \frac{dM}{dH}\right) > 0$$

(provided the tax savings induced by increased medical expenditures exceed the price of medical services), and $dT'/dH < 0$, $dS/dH < 0$

REFERENCES

A. Atkinson and Joseph Stiglitz, "The Design of Tax Structure: Direct Versus Indirect Taxation," *J. Publ. Econ.*, 1976, *6*, 55–75.

William Baumol and David Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

Michael Boskin, "Taxation, Saving and the Rate of Interest," *J. Publ. Econ.*, forthcoming, 1977.

———, "Estate Taxation and Charitable Bequests," *J. Publ. Econ.*, 1976a.

———, "Notes on the Tax Treatment of Human Capital," *Treasury Conference on Tax Research*, forthcoming, 1976b.

——— **and Martin Feldstein,** "The Effects of the Charitable Deduction on Contributions by Low and Middle Income Households," *Rev. Econ. Statist.*, forthcoming, 1977.

Peter Diamond and James Mirrlees, "Optimal Taxation and Public Production: I and II," *Amer. Econ. Rev.*, March and June 1971, *61*, 8–27, 261–78.

Martin Feldstein, "The Effects of the Charitable Deduction on Contributions: I. and II.," *Nat. Tax J.*, 1975a

———, "The Theory of Tax Expenditures," Harvard Institute of Economic Research Discussion Paper, 1975b.

———, "The Welfare Costs of Health Insurance," *J. Publ. Econ.*, 1974a.

———, "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *J. Publ. Econ.*, 1974b.

Arnold C. Harberger, "The Measurement of Waste," *Amer. Econ. Rev. Proc.*, May 1964, *54*, 58–85.

James Heckman, "Shadow Prices, Market Wages and Female Labor Supply," *Econometrica*, 1974

Paul Samuelson, "Tax Deductability of Economic Depreciation to Insure Invariant Valuations," *J. Publ. Econ.*, 1964, *72*, 604–06.

Joseph Stiglitz, "Utilitarianism and Horizontal Equity: The Case for Random Taxation," *IMSSS* Technical Report No. 214, Stanford University 1976a.

Investment and Pricing Policy in the French Public Sector

By H. LEVY-LAMBERT*

The French public sector has become very extensive as the result of post-World War II nationalizations. This situation naturally led French economists to study with special care the methods of investment and pricing to be used by public bodies.

These studies first took place in the electric utility company *(EDF)*. Specific tools were derived and applied to this sector during the 1950's (G. Bessiere and G Morlat, ed.; English text, Nelson, ed.) We will not elaborate on this point, which is well known (for recent developments on optimal control and marginal cost pricing applied to electric utilities, see A. Breton and F. Falgarone, Balasko).

During the 1960's, these tools were tentatively applied to other fields such as intercity transportation, town planning, housing, farming, water resources, etc. This extension was part of a large effort towards improvement of management methods in government, called Rationalisation des choix budgétaires *(RCB)*. *RCB* is somewhat similar to Planning Program Budgeting System. Its development is placed under supervision of Direction de la Prévision (ministry of finance) which supplies technical assistance when required (Levy-Lambert and H. Guillaume).

Criteria were originally based on the classical concept of economic optimum where market prices exist. They were progressively improved in three ways: taking into account imperfection of the environment led to substitute comparison of various possibilities through simulation models to straightforward optimization; taking into account nonmarket costs and benefits led either to tentative valuation thereof or to use of multicriteria decision-making tools; taking into account macroeconomic effects of microeconomic decisions under study led to introduction of shadow prices and addition of special terms to the conventional decision criteria. (See for instance Levy-Lambert and J. P. Dupuy, Vol. 2, Ch. VII.)

These developments were based on works by M. Allais, M Boiteux, J. Lesourne, E. Malinvaud, P. Massé, in microeconomic theory A fairly good review of these works is to be found in J. Dreze. Linkage between micro and macroeconomic tools was analyzed by *CEPREMAP* (R. Guesnerie and P. Malgrange). They derived an endogenous government utility function through analysis of the 6th plan decision process, using a macroeconomic model called *FIFI* (M. Aglietta and R. Courbis). Their work led to a valuation of tradeoffs between household consumption and the main parameters that characterize a given state of the economy, such as unemployment, price level, trade balance, budget deficit or surplus, etc. This valuation was tentatively used as a substitute for surplus maximization as explained below (Section III, Farming).

Long-term models were also used to derive a national discount rate based on the marginal productivity of capital (L. Stoleru, Guillaume). The influence of recent disturbances in the world economy led to its reduction from 10 to 9 percent in real terms (A. Bernard).

The following examples in the areas of transportation, housing, farming and water resources show the way the theoretical tools have been effectively put to use in France

I. Transportation

A) Investment decision making in this area has long been relying on the discounted cash flow criterion, especially as regards roads

*Deputy Manager, Societe Generale (France)

(Direction des routes et de la Circulation Routiere). The benefits computation is mostly based on valuation of nonmarket advantages such as time (M. Giroux 1972) and life (C. Abraham and J. Thedie, Guillaume 1971). Actual values in use are respectively $4 per hour for private cars and $85,000 per life.

Proposed construction of a new high speed railway between Paris and Lyons gave an opportunity for a thorough use of these methods (A. Aurignac). The study included traffic allocation between competitive means, investment and operating costs and pricing problems.

In the field of urban planning, research was undertaken to allow for nonmeasurable objectives such as town development control or users' comfort. Taking these into account, effective ordering of proposed investments was realized for the 6th Plan (1970–75) in the Paris area by means of multicriteria analysis (M. Barbier).

B) Less numerous are applications of economic theory to transportation pricing. Although the theory of variable quality services is well known (Levy-Lambert 1968), its application to roads is postponed by financial considerations While the United States was deciding to progressively suppress tolls on intercity highways, France was conceding the largest part of its proposed highway system for 30 years to private companies as toll roads Conversely, urban highways are free, although they carry heavy traffic. The government intends to charge for them, too, but this step is probably not based on optimal resource allocation considerations.

Several decades after its application to electricity, the domestic airline and the national railroad company decided to try some kind of marginal cost pricing: the days of the year are divided into three price groups (blue, white, red) with half-rate charged during 230 days and a 50 percent premium during 30 days (peak). These experiments seem satisfactory and will probably be extended (C. Azieres).

C) Parking meters were recently installed in Paris and other French cities. Simulation studies had been conducted to estimate the effects of this measure on traffic conditions and on the various groups concerned (Giroux and B. Mourre, Levy-Lambert).

Figure 1 shows, for the case where there is only one means of transportation, the equilibrium traffic when there is no charge and the social optimum. The economic loss due to congestion is also shown. Competition between public and private transportation is described in Figure 2, using allocation curves.

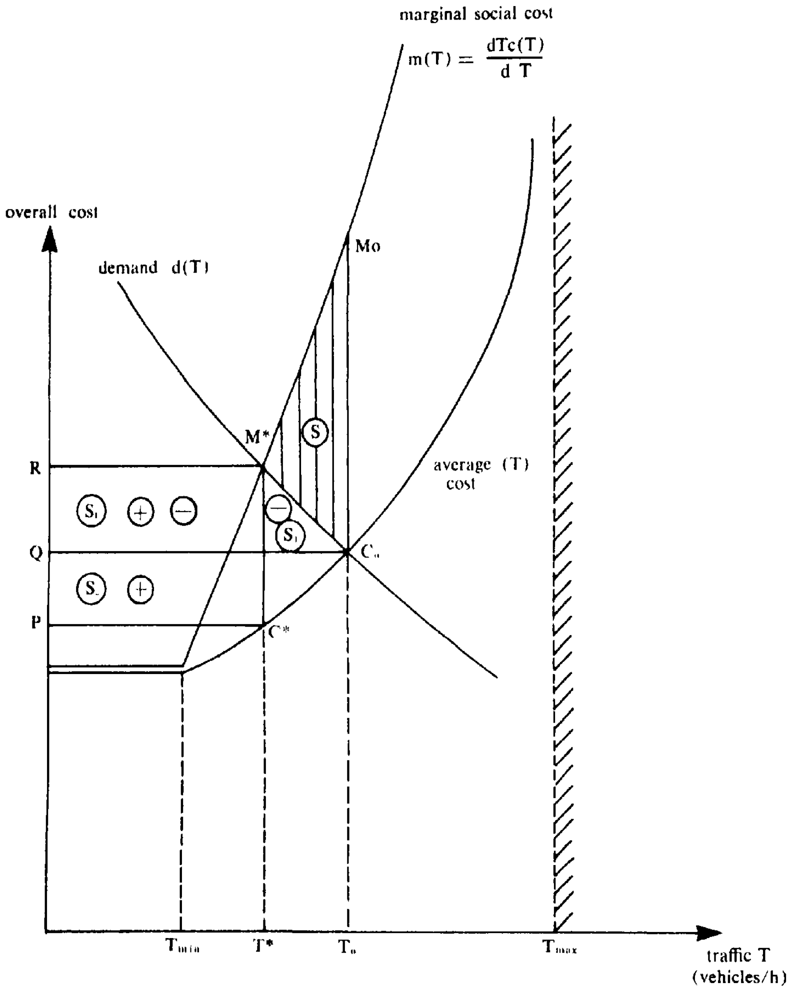
Altogether, the study made showed that moderate pricing of private automobile parking in the streets (50 cents per hour in the center, 25 in the outskirts) would reduce private traffic by 24 percent, increase peak hour speed from six to nine miles per hour, and create a surplus of about $163 million in 1975, most of it being the money value of 600,000 hours of transportation time to be saved per day (Table 1).

D) In order to compare various possible transportation policies in the most extensive way, Direction de la Prévision recently built a simulation model of the whole sector *(BERLIOZ, CITERNE, ROSSIGNOL)*. This model analyzes competition between different means of transportation of goods and people, taking into account investments and pricing policy.

II. Housing

Enormous amounts of public money are spent in the housing sector in France, either by financing or subsidizing certain types of houses or by directly helping certain types of households. Economic analysis is very complex due to market segmentation: similar dwellings will be put on the market at prices very different from one another according to the public subsidies their construction eventually received; also rents paid for a given dwelling can vary according to the subsidies its tenant will receive.

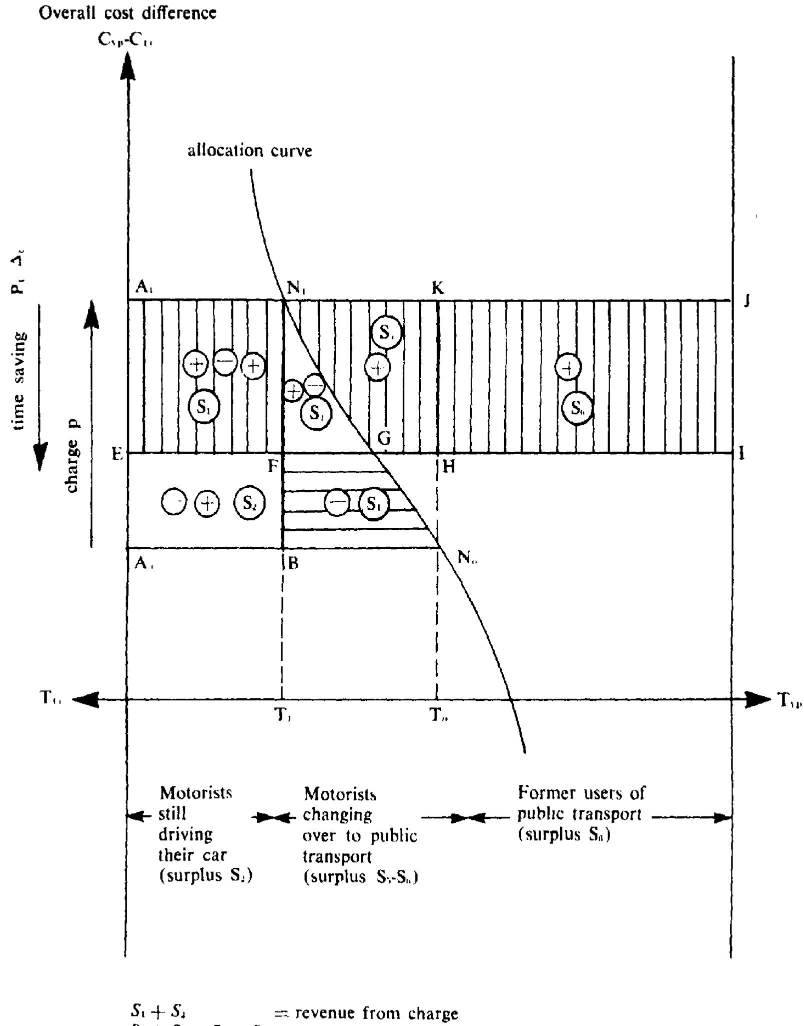
If one wants to assist decision making in this sector, one has to be able to forecast future household choices and first of all to understand their actual choices. This can best be attained through a simulation model such as *POLO*.



T_0 = equilibrium traffic without charge
T^* = optimal traffic (with charge equal to C^*M^*)
$S_1 + S_2$ = revenue from charge
$S_1 - S_3$ = diminution of user surplus

$S_2 - S_3 = S$ = total surplus (user surplus + charge)

FIGURE 1. THE GENERAL CASE



$S_1 + S_2$ = revenue from charge
$S_1 + S_2 + S_0 - S_1$ = total surplus (user surplus + charge)

FIGURE 2. COMPETITION BETWEEN PRIVATE CARS AND PUBLIC TRANSPORT

TABLE 1—BALANCE OF GAINS AND LOSSES AS A RESULT OF CHANGEOVER FROM 9 km/h TO 14 km/h IN 1975.

Group Concerned	Trips per day (000)	Gains (+) and losses (−) (million francs per year)			
		Cash	Time	Convenience	Total
1. Former users of public transport	558		+162		+162
2 (a) Motorists changing over to public transport	613	+25	+66	−277	−186
(b) New bus users other than (2a)	379		+42		+42
3. Motorists still driving their own car					
(a) Going to Paris and paying the parking charge	1,472	−567	+393		−174
(b) Going to Paris to free parking facilities	108		+54		+54
(c) Going to commuter belt	347		+133		+133
4. Local authorities (net revenue from parking charges)		+478			+478
5. Transport enterprises		+307			+307
Total		+243	+850	−277	+816

The *POLO* model has been devised by Direction de la Prévision. It helps decision making in housing aids as well as in investment both in quantity and in quality (Levy-Lambert 1968, Y. Carsalade and R Pincon, 1973) See Figure 3 Allocation between housing x and other goods q is based on maximization of a very simple utility function. Volume of consumption of housing goods is defined by the actual construction cost of a dwelling equivalent to the one used. Allocation problem is the following, p standing for the consumption price index and 1 for the annual cost of a house (in %):

$$S(q, x) = q^{1-a} x^{a} \text{ max!}$$

$$pq + 1x = r$$

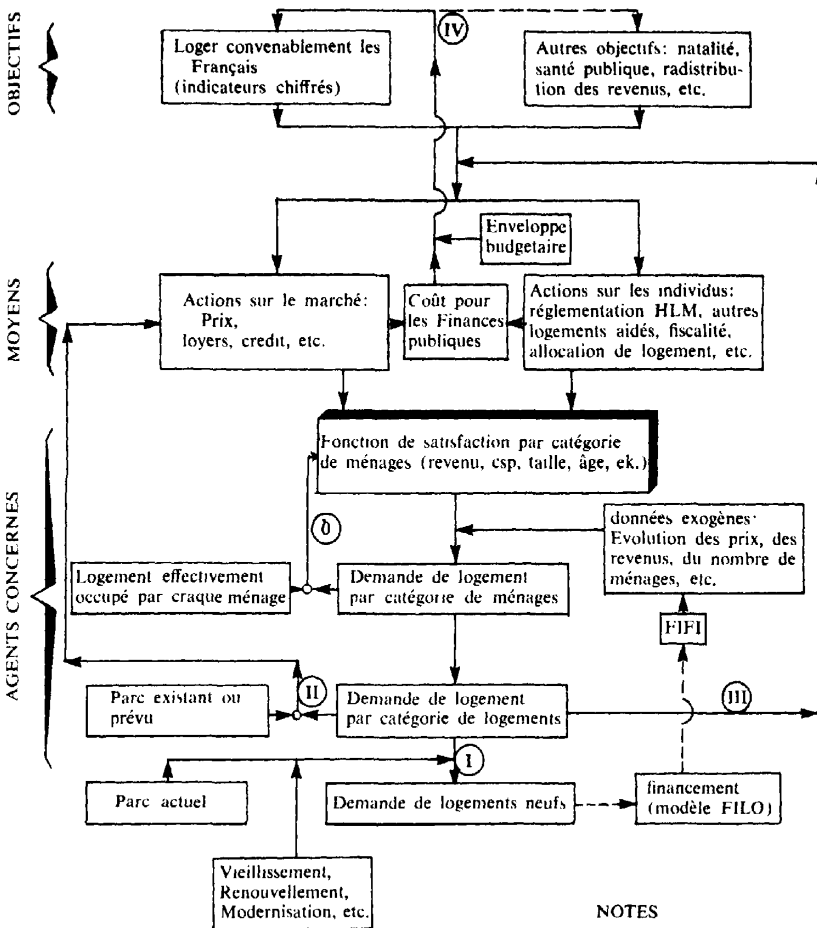
Knowing the dwelling each household has actually chosen amongst all those which he could theoretically select (l_1, x_1 − l_2, x_2 − l_3, x_3.) gives way to determination of the most likely values for the set of a parameters for each household. This can be done for several past years and helps in estimating future values through multiple regressions.

Clearly if there were a continuous set of x values and only one value of l, parameter a would simply be the proportion of income each household is devoting to his home. Segmentation and differentiation of markets make it somewhat more difficult to determine values of a. Yet this was done by using *INSEE* (National Statistics Institute) periodical investigations of housing conditions. a values ranged in 1967 from less than 10 percent for 26 percent of the population to more than 30 percent for 19 percent of the population. The average value was about 20 percent.

Knowing the value of a for each household allows the next step, which is simulation of various housing aid policies at a given time. Each policy can be specified by values of l_{ij} associated with each type of dwelling x_i and each type of household j (subscript i stands for so-called "stone aids" and subscript j for so-called "personal aids"). Policies studied included free rent, deletion of stone aids, various formulas of personal aids, etc.

Incidentally, these studies gave confirmation of personal aids being the cheapest way for public finance to attain a given housing objective.

POLITIQUE DU LOGEMENT (POLO)



NOTES

0 Adjustment of the parameters of the model (done once and for all).

I Determination of a building program.

II The demand is set equal to supply by price adjustment.

III Change of policy instruments to accommodate the demand and its impact on the goals.

IV Adjustment of goals in view of budgetary constraints.

FIGURE 3. HOUSING POLICY

If the governmental housing policy is now fixed, the *POLO* model tells how demand for each type of dwelling will vary with time, knowing the variation of the number of households, their income and their parameter a, provided depreciation policy is known. Con-

versely, knowing the number of dwellings of each type existing at a given time allows one to determine equilibrium prices, that is, prices allowing supply and demand equality for each type. Hence, simultaneous determination of optimal construction and depreciation policy follows, the construction cost of any dwelling having to be equal to the discounted value of its future incomes. T being the period of time during which a type of dwelling is effectively constructed, $l(t)$ its rental rate and $e(t)$ its operating costs, one can write (Cf. for example, Levy-Lambert and Dupuy, Vol 1, p. 127):

$$\int_{t_c}^{\infty} [l(t) - e(t)]\exp - it\, dt = 1 \quad \forall\, t_c \in T$$

Hence, by derivation:

$$l(t) = i + e(t) \qquad \forall\, t_c \in T$$

For $t_c \notin T$, the number of dwellings of the type under consideration being constant, $l(t)$ is given by equalizing supply and demand. The case when $l(t) < e(t)$ means this type is to be put out of use

Actual computations show, as could be expected, that allowance for future increase of incomes results in higher average quality of construction programs than would result from actual demand.

III. Farming

A) Agriculture is heavily subsidized in every country. France is no exception. Subsidies are of many kinds, including investment grants, social funds, tax reductions, direct subsidies, price support, etc.

Price support is the most popular means among farmers since it seems to ensure them a minimum income. It is indeed the most expensive means for public finance since it is more profitable to the wealthy. It is also the most inefficient because of its side effects of increasing supply and decreasing demand, as long as prices are artificially maintained above the average equilibrium level. Moreover its real cost is higher than appears since part of it is directly paid by the consumers.

Straightforward computation of economic surplus indicates economic loss due to price support. It can be reduced by means of deficiency payment but the cost for public finance increases rapidly.

B) The above conventional approach is static and does not take into account the effects of policies studied on concerned groups and on the economy as a whole. The following study is a step towards a more inclusive approach of public decision making in the agricultural sector. Focus was set on production and employment. Discounted benefits and costs of increasing the natural rate of reduction of the agricultural population were computed (Levy-Lambert and J C Papoz)

Table 2 shows the results of the computations The reference solution is a straightforward continuation of past tendencies with population reduced by one half (3,050,000 to 1,560,000) from 1967 to 1985 *(A)*. Three other possibilities were studied, leading to final populations of respectively, 1,330,000, 830,000 and 500,000 *(B, C, D)*.

Lines 1 to 4 indicate effects on the agricultural sector itself (consumption, investment, stocks). The following lines deal with other effects The change in foreign trade due to the agricultural production decrease is on line 5 (internal consumption is assumed to be unaffected) Production changes are of course valued at international market prices when different from internal prices. Costs of job conversion appear on line 6, including training, housing and job creation. The benefits of job conversion, based on gross added value per capita, are on line 7, assuming converted farmers actually find a job.

Summarizing, Table 2 shows the high social profitability of the policy studied. Pay-out time is 5 years This result simply stands for the fact that the marginal productivity of labor is lower in agriculture than in other sectors.

The surplus computation above does not allow for side effects nor does it take into account

TABLE 2—SUMMARY OF COSTS AND BENEFITS OF ALTERNATIVE SHIFTS OF THE AGRICULTURAL STRUCTURES
(in millions of 1965 Francs, 1975 prices, real value)

	1975			Total 1970-85		
	B - A	C - A	D - A	B - A	C - A	D - A
1. Intermediate Consumption	245	1,355	2,320	5,880	32,515	55,675
2 Gross Fixed Capital Formation in Agricultural Sector	517	1,190	1,527	4,005	11,052	13,786
3. Cost of Structural Shifts	−108	−245	−262	—	—	—
4 Final Inventory	—	—	—	−1,180	−4,930	−8,065
5. Foreign Trade	−115	−1,136	−2,770	−2,775	−32,760	−66,480
6. Conversion Cost	−1,072	−3,429	−5,038	−8,342	−26,676	−39,176
7 Conversion Benefits	1,265	4,048	5,945	45,790	146,529	215,213
Total not discounted	732	1,554	1,722	43,298	125,730	171,043
Total discounted at 10% to 1970 Base				13,438	36,769	48,540

government valuation of other objectives than final consumption.

Guillaume (1973) tried to compute the relevant corrections using results of M. Aglietta and R Courbis FIFI macroeconomic model and of Guesnerie and P. Malgrange government utility function. Computation was made only for 1975, the last year of the 6th Plan.

Table 3 gives results of the computation made with one set of weights. These results are quite different from those of the partial analysis above. Three reasons account for this departure:

1) The increase in household consumption according to *FIFI* is smaller. This is mostly due to the fact that part of the farmers become jobless with *FIFI*;

TABLE 3—MACROECONOMIC IMPACT OF ALTERNATIVE SHIFTS OF THE AGRICULTURAL STRUCTURES
(Year 1975)

	Weight in the *SPF*	B - A	C - A	D - A
Gross Domestic Production (*GDP*) in Quantity	—	353	3,104	4,736
Growth Rate of *GDP* (in %)	—	+ 01	+ 04	+0 07
Production of the Agricultural Sector in Quantity	—	−703	−3,517	−7,034
Final Capital	1	338	1,510	3,930
Number of People Looking for a Job	−15	+13	+31	+46
Number of People Being Transferred	11	−58	−182	−266
Financing From Nonagricultural Sectors	1,1	−422	−64	−602
Household Consumption	1	+431	+851	+1,314
Rate of Change of the General Price Level	−500	—	−0.03	−0.05
Rate of Growth of Per Capita Wages	—	—	−0 04	−0.06
Financing by Government	0.6	−546	−2,546	−5,199
Fiscal Burden	3,000	—	—	—
Collective Consumption	1.36	+105	0 15	0.25
Rate of Growth of Industrial Production (in % per year)	500	0.05	0 15	0.25
Change in the State Preference Function		−688	−1,084	+116
Pure Microeconomic Benefits Reported from Table 2		+732	+1,554	+1,722

2) Other terms in the government utility function stand for a governmental reluctance to any important change, e.g., in terms of jobs;

3) This macroeconomic approach is static and cannot take into account changes with time that appear in the discounted cost benefit approach.

Yet this experience proved useful by its tentative reconciliation between micro and macroeconomic approaches.

The above study was mechanical and led to rather tedious manual calculations. To study numerous more sophisticated possibilities, J. M. Ruch, A. Monfort and G. Winter (Direction de la Prévision) devised a long-term simulation model of the agricultural sector. This model determines population, investments and production using production prices as exogenous variables (the breakdown of products is into 7 groups). This model helps in analyzing various policies as regards aids, production, financing, etc.

IV. Water Resources

In France as in many other countries, water resources legislation was long based on standards derived from equity considerations rather than from optimal resource allocation principles. The basis for a more rational system was set in 1964 (Allen V. Kneese, pp. 269-274) when six *basin agencies* were created These agencies cover the whole country according to river basin limits. They charge all water users and polluters including farmers, industries and communities and help finance public and private investments.

A theoretical basis for the price system may be found in Levy-Lambert (1964). It is a straightforward application of social marginal cost pricing: rates for use of water or discharge at any place have to be equal to the sum of all related marginal costs imposed to users downstream. When such costs cannot be measured, they are replaced by standards and the relevant rate is equal to the associated dual variable. Such cases arise with the salt content in drinking water or with river temperature.

When there are large collective works for water transfer or storage or waste treatment, their marginal operating costs supply a basis for the determination of the rate structure. Of course, actual rates implemented by basin agencies do not exactly comply with theoretical formulae, due to difficulties of estimating the relevant data as well as political reasons. The latter lead to a progressive move towards optimal rates.

Rates were nevertheless fixed so that their relative variations would be as close as possible to variations of the economic value of water with respect to time, space and quality. Therefore, river water use rates are generally low or zero during the wet season and high during the summer. Also discharge rates are usually higher upstream (unless there is sufficient self-purification) than near the mouth (the ratio is about ⅓ in the Seine basin) Rates applying to waste content of effluents are related to separation costs of each type of waste in treatment plants

The French experience of water resources pricing is unique in its extension throughout the country. Its effects on improving the environment are already noticeable.

REFERENCES

C. Abraham and J. Thedie, "Le prix d'une vie humaine dans les décisions économiques" (The price of human life in economic decision making), *Revue francaise de recherche opérationelle, 16,* 3ème trim. 1960.

——— **and A. Thomas,** *Microeconomics, Optimal Decision Making By Private Firms and Public Authorities.* Translation by D. V. Jones. Reidel, 1973.

M. Aglietta and R. Courbis, "Un outil pour la Plan: le modèle FIFI" ("A tool for the Plan: the FIFI model"), *Economie et Statistique, 1,* mai 1969, 45–65.

M. Allais, *La gestion des houillères nationalisées et la théorie économique (The*

management of nationalized coal mines and economic theory), Paris 1953.

A. Aurignac, "Etude des transports rapides sur l'axe Paris-Sud-Est" ("Study of rapid transport, Paris-Southeast"), *Bulletin RCB*, 7 mars 1974, 27–41.

C. Azieres, "La SNCF et les vacances" ("The SNCF[1] and vacations") *Bulletin du PCM*, Sept. 1976.

Y. Balasko, "Formes optimales de la tarification de l'électricité" ("Optimal tariffs for electric utilities"), *UNIPEDE* Conference, Madrid 1975.

M. Barbier, "La rationalisation du choix des investissements de transport pour le VIè Plan en Région Parisienne" ("Investment decision for transport in the Paris Region under the Sixth Plan"), *Bulletin RCB*, 4 juin 1971, 22–31.

J. Benard, "Les modèles d'optimation économique de l'éducation" ("Models for economical optimal education"), *Revue d'économie politique*, 3 mai-juin 1973, 433–481.

C. Berlioz, P. Citerne and J. P. Rossignol, "Le modèle de simulation de politiques des transports" ("A model for simulation of transport policies"), *Statistiques et études financières*, série orange, *23*, 1976, 3–19.

A. Bernard, "Une nouvelle évaluation du taux d'actualisation pour l'économie francaise" ("A new estimate of the discount rate for the French economy"), *Revue économique*, 3 mai 1972.

______, "Le taux d'actualisation du Plan et la crise de l'énergie, colloque franco-japonais sur la planification" ("The discount rate of the Plan and the energy crisis—French-Japanese colloquium on Planning"), Tokyo 1976.

G. Bessiere and G. Morlat, eds. *Vingt cinq ans d'économie électrique (Twenty-five years of electricity economics)*, Dunod 1971.

D. Blain and P. Morin, "Hausse du prix du pétrole et choix économiques" ("Increase in oil prices and economic choices"), *Statistiques et études financières*, série orange, *21*, 1976, 27–64.

M. Boiteux, "Sur la gestion des monopoles publics astreints à l'équilibre budgétaire" ("On management of public monopolies constrained to balanced budget"), *J. Econ. Theory*, 1972 (translated).

C. Bozon and C. Charmeil, "La programmation des investissements routiers sur une liason" ("Investment in highway systems"), *Revue générale des routes et des aérodromes*, 395, janvier 1965, 119–41.

A. Breton and F. Falgarone, "Application de la théorie de la commande optimale au problème du choix des équipments de production à Electricité de France[2]" ("Application of optimal control theory to fixed investment selection at 'Electricité de France'"), Fourth Power Systems Computation Conference, Grenoble 1972.

______ **and M. Cremieux,** "La séparabilité dans le temps du choix des équipements de production à E.D.F." ("Intertemporal separability of fixed investment selection at 'Electricite de France'"), ORSA, TIMS conférence, Philadelphia 1976.

Y. Carsalade and R. Pincon, "POLO: modèle de politique du logement" ("POLO: a model for housing policy"), *Statistiques et études financières*, série orange, *10*, 1973, 17–52.

R. Cros, "L'optimum de second rang et la tarification des entreprises publiques" ("Second best pricing policy for public utilities"), *Revue d'économie politique*, *6*, nov.-dec. 1973, 986–1021.

J. Dreze, "Some Postwar Contributions of French Economists to Theory and Public Policy," *Amer. Econ. Rev.*, June 1964, *54*, 1–64.

M. Giroux, *La tarification des transports intérieurs (Pricing of domestic transportation)*, Bordas, Paris 1973.

______, "La valeur économique du temps et

[1]State owned railroad monopoly.

[2]State owned monopoly supplying electricity.

les transports" ("Economic value of time and transportation"), *Bulletin RCB*, 8 juin 1972, 27–35.

——— **and B. Mourre,** "Le stationnement payant" ("Paying for parking"), *Statistiques et études financières*, série orange, *5*, 1972, 3–14.

R. Guesnerie, "Un formalisme général pour le second rang et son application à la définition des règles du calcul économique public" ("A general framework of the second best theory and its application to public economics"), *Cahiers du séminaire d'économie C.N.R.S.*, Paris, *16*, 1975, 87–116.

——— **and P. Malgrange,** "Formalisation des objectifs à moyen terme. Application au VIè Plan" ("Formalization of intermediate run targets. Application to the 6th Plan"), *Revue économique*, 3 mai 1972.

H. Guillaume, "L'étude économique du Réseau Express Régional" ("Economic study of the Réseau Express Régional"[3]), *Calcul économique II*, PUF, Paris 1971.

———, *Prix fictifs et calcul économique public (Shadow prices in public economics)*, C.N.R.S., Paris 1973.

———, "Le coût économique de la vie humaine" ("The economic cost of human life"), *Bulletin RCB*, n° 5, sept 1971, 45–58.

——— **and P. Rochard,** "Compatibilité entre approche sectorielle et globale" ("Compatibility between sector and aggregate approach"), *Statistiques et études financières*, série orange, *9*, 1973, 25–59.

M. Guillaume, "L'évaluation du taux d'actualisation associé à la croissance francaise" ("Estimate of the discount rate associated with French economic growth"), *Economie appliquée*, 1968, *34*, 917–59.

P. d'Iribarne, "Valeur de la vie humaine et politique rationnelle de santé et de sécurité" ("Value of human life and rational health and safety policy"), *Analyse et Prévision*, *6*, décembre 1969, 725–36.

A. V. Kneese and B. Bower, *Managing Water Quality*, Baltimore 1968.

S. C. Kolm, "L'état et le système des prix" *(The state and the price system)*, Dunod, Paris 1971.

L. Lebart, *Recherches sur le coût de protection de la vie humaine dans le domaine médical (A study on the cost of protecting human life in the health sector)*, Crédoc, Paris 1970, multigr.

J. Lesourne, *Technique économique et gestion industrielle, (Economic analysis and industrial management)*, Dunod, Paris 1972.

———, *Le calcul économique (Economic Calculus)*, Dunod, Paris 1965.

V. Levy-Garboua, "L'analyse de surplus appliquée" ("Analysis of surplus"), *Revue économique*, *6*, nov. 1975, 987–1003.

H. Levy-Lambert, "L'eau, abondance ou pénurie: une question d'organisation" ("Water, abundance or scarcity: a matter of organization"), *Annales des mines*, sept. 1964, 579–604.

———, "Tarification des services à qualité variable: application aux péages de circulation" ("Pricing of services with quality fluctuations; application to traffic tolls"), *Econometrica*, *36*, juillet 1968, 564–74.

———, "Modèle de choix en matière de politique du logement" ("Model of choice with respect to housing policy"), *Revue d'économie politique*, *6*, 1968, 937–67.

——— **and J. P. Dupuy,** *Le calcul économique dans l'entreprise et dans l'administration*. Tome 1: principes de base—Tome 2: études de cas (*Economic calculus in firms and government*. Tome 1: principles—Tome 2: case studies), Dunod, Paris 2e ed. 1975.

——— **and H. Guillaume,** *La rationalisation des choix budgétaires (Rationalization of budget choices)*, PUF, Paris 2ème éd. 1975.

——— **and J. C. Papoz,** "Coûts et avantages pour la Nation de l'évolution des structures agricoles" ("Social cost and benefits of agricultural transformation"), *Economie*

[3]New network of train transportation for commuters in the Paris area

rurale, *86*, 4ème trim. 1970, 51–68.

——, "Cost Benefit Analysis and Urban Traffic Congestion; The Example of Paris." *Transport and the Urban Environment*, J. Rothenberg, ed., New York 1964, 223–39.

E. Malinvaud, *Lecons de théorie micro-économique (Lectures on microeconomic theory)*, Dunod, Paris 1969 (Trans. New York, 1972).

—— **and Y. Younes,** "Nouvelle formulation générale pour l'étude des fondements microéconomiques de la macroéconomie" ("New general framework for studying the microeconomic foundation of macroeconomics"), *CEPREMAP*, 1974, miméo.

P. Masse, *Optimal Investment Decisions: Rules for Action and Criteria for Choice*, Translated by Scripta Technica, Inc., Englewood Cliffs 1962.

J. M. Ruch, A. Monfort and G. Winter, "Un modèle agricole à long terme de simulation" ("A long term simulation model of agriculture"), *Statistiques et études financières*, série orange, *16*, 1974, 27–51

L. Stoleru, "Répartition des investissements et taux d'intérêt" ("Allocation of investments and interest rates"), *Revue d'économie politique*, *6*, nov./dec 1967, 829–47.

G. Terny, *Economie des services collectifs et de la dépense publique (Economics of public goods and public expenditures)*, Dunod, Paris 1974.

Commissariat General au Plan, *Calcul économique et planification (Economic calculus and planning)*, Doc. Frse, Paris 1973.

Direction des Routes et de la Circulation Routiere, "Calculs de rentabilité appliqués aux investissements routiers" ("Profitability analysis applied to highway investments"), Ministère de l'Equipement et du Logement, Paris 1970 (révisé en 1974).

Institut D'Amenagement et D'Urbanisme de la Region Parisienne, "Les transports urbains et leurs usagers dans la région parisienne: choix de moyen de transport par les usagers" ("Urban transportation in the Paris area; users choice of the transportation modes"), *Cahiers de l'I.A.U.R.P.*, 1966, *2*.

——, "Analyse du choix du mode de transport par les usagers en région parisienne" ("Users choice of transportation mode in the Paris region"), *Cahiers de l'I.A.U.R.P.*, 1969, *2*.

Service des Affaires Economiques et Internationales, "Travaux à long terme en matière de transports" ("Long term studies in transportation"), *Objectif*, numéro spécial, juin 1970, Ministère de l'Equipement et du Logement, Paris.

Discussion

DAVID BRADFORD, Department of the Treasury: Let me begin by noting two points of comparison between the Hubert Levy-Lambert paper and the Joseph Stiglitz-Michael Boskin paper. First, in spite of the apparent concern with the expenditure side of the government in the Levy-Lambert paper and the tax side in the Stiglitz-Boskin paper, they have in common that they are concerned with issues of pricing. Many of the decisions involved in transportation, housing, agricultural policy and water resources described by Levy-Lambert are pricing decisions. Tax policy, of course, also concerns the appropriate relationship among prices and costs. The second point concerns a contrast between the two papers. The Levy-Lambert paper illustrates the widespread use of economic models and economic theory in making public policy decisions throughout the French government. The Boskin-Stiglitz paper, on the other hand, is more speculative, and I think this correctly reflects the difference between France and the United States in the degree of application of economic analysis.

What have been the "recent developments" reflected in the application described in these papers? By and large these have been an evolution, rather than a revolution, with the common elements of increasing sophistication, in dealing with the second best nature of public policy problems, in attempting to make quantitative estimates of important parameters of the problems, and in elaborate modeling. While these papers show that a great deal of good work has been done, they have led me to consider some of the gaps that remain. Let me discuss a few of these in turn.

One of the problems which needs to be dealt with in an analysis of a public policy choice is identifying the "real" constraints. Second-best theory has increased our awareness of how important it can be to identify these and has shown how the results may be affected by them. Unfortunately, our policy recommendations will often depend as much on what we assume to be the constraints as on, for example, estimates of elasticities. Some examples:

1) In designing programs to raise employment of central city young people, is it a constraint that no employer should be allowed to have a labor cost below the federal minimum wage?

2) In considering the use of central city housing, is the central city taxing and expenditure system to be regarded as immutable?

3) In approaching the problem of efficient allocation of population to the agricultural sector, as described by Levy-Lambert, which of the existing policies which presumably caused the difficulty are to be regarded as fixed?

4) In considering the desirable tax treatment of capital gains, is inflation indexing to be regarded as impractical?

5) Considering the question of the desirable way to tax corporation income, is the existing method of taxing capital gains to be regarded as fixed?

6) Considering the construction of a dam which will have as a by-product recreation potential, should one regard as fixed an inefficient policy of pricing the resulting facilities?

7) Considering the regulation of telecommunications, must we assume that rate of return regulation is the policy by which the monopolistic segment will be controlled?

A second and probably more serious problem is the lack of knowledge about the correct models. That is, in many cases, we are not confronting the difficulty of estimating parameters of a reasonably well understood structure, but of groping with quite badly understood systems.

1) In the tax field, this is illustrated by the continuing controversy over the incidence of the corporation income tax. This argument implicitly concerns the appropriateness of the competitive model to describe the allocation of a large portion of the nation's investment resources, so that the argument could be seen as

implicitly about the validity of a great deal of our modeling of the effect of taxes.

2) In regulation of telecommunications, the role of competition as an innovative force is still not thoroughly understood. Thus the theory of optimal taxation would presumably tell us to extend the monopolistic sector as far as possible, so as to make as innocuous as possible the budget balancing constraint. What is the nature of the offsetting gains to be had from allowing competition in some portion of the industry?

3) What will the reaction of political systems be to changes in the rules, for example, for revenue sharing or for the tax treatment of local taxes, or to court decisions on racial integration in schools?

4) What determines the level of capital stock demanded by individuals; what is the effect on private demand of anticipated future taxes (e.g., to retire debt, to pay for social security, etc)?

5) What is the effect on the equilibrium distribution of wealth of higher or lower inheritance (or estate) taxes? What is the effect on the distribution of leisure, consumption, etc., of any of these?

While on each of these issues we can point to work which has been done, in none of them is there widespread agreement about the correct model.

At a more philosophical level, there are difficulties with applying economic theory arising from the lack of well defined objectives.

1) What are the proper weights in those welfare functions? Indeed what is the entity which has a utility function? This problem has arisen in an interesting way recently in thinking about the treatment of the family in the tax system. For some purposes it seems reasonable to have the family as the "unit" of welfare, while in others we want to think about the individuals in the family separately. This makes it hard to apply optimal tax theory.

2) Does a welfare function at all describe the way people's choices are ordered in a political framework? Two examples come to mind. The paper cited by Levy-Lambert, by Ginnery and Margams, derives the preferences over a set of objectives implicit in the sixth plan in France. A recent paper by Peter Diamond on the social security system does something similar for that institution. Both of these efforts, and others in the literature, might be regarded as figuring out the problem to which a particular policy is the solution, and generally this does not look like maximization of a welfare function of the conventional sort.

3) Where in the welfare maximization story do we place the notion of "fair burden," which plays such an important role in the political debate about taxes. Where does the notion of "freedom" fit in, which many regard as the true touch stone of policy?

4) Finally, we must ask whether "society's" choices reflect an ordering at all. This is the problem of social choice theory. Even if we were to discover that society's choices are so arrayed, we might well ask whether this preference structure deserved normative power. Certainly there are decisions of some government's which we do regard as wrong. At this level we are led to question how the information provided by economists will be used, much as physical scientists have long concerned themselves with the political uses of advances in their field of knowledge.

The two papers we have heard suggest that progress is indeed being made. One is reminded of Keynes' observation about the power of academic scribblers, admittedly occurring with long and variable lags. Much public debate now seems to me to be carried out with the tools of a decade or two ago and it should be interesting to see how today's "recent developments" make their way into the public discourse a decade hence.

Ethics in Economics

By LEONARD SILK*

Ethics increasingly has made its way onto the agenda of the American economics profession. This has happened, I believe, for two reasons, one exogenous and the other endogenous.

The exogenous cause was the Nixon Administration, and the more conspicuous-than-usual bowing of the knee by economists to the sovereign. That was bad enough; it certainly had its effect on the quality of analysis and forecasting, and the mode of public discourse among economists, which lowered the standing of the economics profession in the eyes of the wider public. The episode was made all the more unpleasant by the personal-cum-political hostility displayed by some Nixonian economists toward their critical or even only disagreeing colleagues in the profession One eminent economist, Paul A. Samuelson, even made it onto the enemies list. Samuelson believes that he made the list by his achievements as a journalist rather than as an economist And it must be added in fairness that the list was the work not of Administration economists but of political operatives, the same operatives who discussed plans to bomb the Brookings Institution. Nevertheless, the mood of the period was poisonous, and I regret to say that some in-house, that is White House, economists were infected by it.

I am delighted to tell you that, in my opinion, that mood has been dispelled and that, across the frontier between the conservative economists who serve the Ford Administration and the liberal economists who do not, peace and a considerable measure of mutual respect reign. I would specifically credit Alan Greenspan, Chairman of the President's Council of Economic Advisers, with a well-thought-out and well-executed plan to restore the credibility of the economists in government. If the economists' standing is still lower than it was during the glory days of the Kennedy and early Johnson years, or year, that is more due to the persistent weaknesses of the economy and the inability of economists to come up with convincing solutions, convincing even to themselves, rather than to a belief that they are dissimulating or faking the evidence At the personal level, relations between highly placed economists within the Administration and outside it have markedly improved. I hope that I am not overstating the case. Insiders will know better than outsiders whether and how much leaning on professional staff for predetermined conclusions has still been going on, and whether there has been any significant massaging of the data. On the data, I firmly believe there has been little or none; and this is doubly commendable because the data, especially on unemployment, have been extremely inconvenient for an incumbent President in this election year. Given uncertainties about the seasonal adjustment factors, this demonstrated restraint and professional rectitude on the part of the President's men—that is, his economists—has been almost above and beyond the call of duty. I hope it continues, and into the next Administration.

I asserted earlier that there was also an endogenous reason, endogenous within the discipline of economics as James Tobin uses the term, why ethics has moved up on the profession's agenda. That reason is the increasing awareness among economists of the necessity of paying more attention to the *goals* of eco-

*The New York Times

nomic policy, and these are value-laden, rich in ethical content and deep in ethical confusion. A growing number of economists now believe that the economist himself should concentrate more on the entire process by which goals are set and on helping to resolve conflicts among different goals, such as freedom, efficiency, equity, security, stability and growth. In the past economists have customarily taken such goals, in their more specific forms, as "given"—that is, given by other policy makers or by the society at large—and have maintained that goals are subjective; hence, that economists are no better than anyone else at setting them. All the honest economist, as economist, could do would be to display alternate means of achieving goals, and point out inconsistencies, when such existed. But this modest or self-limiting posture has led to a very unhelpful vagueness in an area that is crucial to the entire realm of economic and social policy. For instance, the new report on economic education, being prepared for the Joint Council on Economic Education under the chairmanship of W. Lee Hansen, says of the goal called "equity," an extremely important ethical concept, "This is an elusive concept. There is no agreement on what is equitable; people differ in their conception of what represents fairness or equity. In evaluating economic performance, the concept is essential in reminding us to investigate who or what kinds of people are made better or worse off as a result of a change in prices or the implementation of a new government program."

But because the goal of equity is difficult to define, does that imply that some other goal, such as efficiency, should supervene? My own answer—and I should think most of yours—would be no. Yet, in the absence of a better means of dealing with the question of equity, the goal of efficiency may in fact supervene in the work of the economists—and give rise to a sometimes bitter or cynical feeling that others, especially politicians, are "wrong" in rejecting their conclusions.

In the real world of politics, it is the clash over values and goals that is the essence of the policy problem. All economists know this, and it variously informs the work and thought of our most important economists, such as Samuelson, Milton Friedman, John Kenneth Galbraith, Wassily Leontief, and Kenneth Boulding, whose lives and ethical views I have recently sought to explore. I found the exercise, when done in terms of specifics, not generalities, extremely helpful as a means of gaining perspective on the divisions and confusions within contemporary economics.

Within economics, as within the whole of modern Western society, we continue to re-enact the philosophical history and confrontations of ancient Greece. We have our Cynics, descendants of Diogenes, believers in a philosophy of retreat; they hold that life of man in society is bad and one can find satisfaction only in unresisting resignation to the evils of the world. We have our Skeptics, heirs of Pyrrho, who assert that there could never be any rational ground for preferring one course of action over another, and that the task of man is simply to conform to the customs of the country where he lives; a view also held by many businessmen, who now operate in many countries. We have our Epicureans, children of Epicurus, who hold to a philosophy of preferring pleasure to pain; the felicific calculus is indeed at the heart of conventional economics. Yet it is worth reminding ourselves that Epicurus thought true pleasure was to be found in moderation. "The greatest good of all," he wrote, "is prudence; it is a more precious thing than philosophy." Epicurus wanted man to avoid fear and to abstain from public life. But that is advice very few economists are willing or able to accept, once the opportunity of a public life and power is thrust upon them. And many, alas, even connive at achieving the opportunity, as we say, to serve.

But that is not all; we have our Stoics, our Platonists, our Aristotelians, and, coming down to later epochs, our Christians and our Marxists. I offer all of this rich ethical material to you as a gold mine, perhaps a uranium mine, for further exploration.

However, I must move on, since my chairman has also asked me to comment on ethics in business as compared to ethics in government. This, too, is a subject that has been close to my heart and my typewriter in recent years.

Public confidence in the leadership and integrity of business has been gravely weakened in recent years by revelations of the efforts of some corporations illegally or unethically to influence government decisions and policies, both at home and abroad. The cases have been too numerous and important for anyone to dismiss as a fabrication or distortion by the "media." Watergate, the Lockheed-Tanaka connection in Japan, and the Lockheed-Prince Bernhard connection in the Netherlands (as well as the complicating Northrop-Bernhard connection), which have shaken the stability of governments, have been only the most conspicuous and sensational of these business-government scandals. And there are doubtless other revelations to come.

Why have so many businesses entered into corrupt relations with highly placed government officials? The most obvious answer, and one that many members of the press and public may think is sufficient to explain the whole story, is that certain businessmen (with the cooperation, encouragement or even extortionate pressure of government and political leaders—it does take two to tango) have put immediate corporate sales and profits and their own personal interests above all other considerations, including respect for the law. On a straight cost-benefit analysis, such corporate officials decided that "corruption pays."

The thesis that corruption pays certainly does not apply to all business situations or, I would maintain, the vast majority of them. It was Adam Smith—the moral philosopher whom we honor this year for the 200th anniversary of *The Wealth of Nations*, together with our nation's bicentennial—who pointed out that among those who trade often with each other honesty is the best policy. That is still normally true for relations among investors and their brokers, purchasers and suppliers, bankers and their customers, and many others who constantly do business together: honesty and integrity are consistent with long-run stability of relationships and long-run profit for firms. Honesty is clearly a "public good" for the participants in certain markets, since all benefit by preserving a code of fair dealing and free competition. Capitalism was believed to rest on a moral foundation.

Then what has gone wrong—and why has corruption become so serious a problem in the United States and elsewhere? Among the possible explanations are these:

1) There has been decay of traditional morality in the society at large. A generation ago, in his *A Preface to Morals*, Walter Lippmann said that virtue was not the creation or monopoly of the tender-minded and sentimental but derived originally from a profound realization of the character of human life; and that widespread social and personal immorality was due directly to the loss of genuine belief in the premises of popular religion. This loss of belief is now further advanced

2) The growing anonymity of life in mass society; the lack of close personal relations between corporate leaders and other members of the community, representing different interests and different points of view; the huge size of firms, their bureaucratic structure, the loss of a sense of family tradition and honor as firms increasingly become "public" corporations, the short-run perspective imposed not only by immediate market pressures but by stockmarket considerations, the diminishing importance of a reputation for "character" in the making of money—all such factors add up to a diminution of the force of social (as opposed to governmental) controls

3) Corruption has been growing, *pari passu*, with the growing weight of government in the market, and the importance to businessmen of eliciting government actions or policies favorable to particular corporations. This is particularly true in the defense area, but it goes far beyond military procurement in an age in which contracts or regulation or subsidy or licenses or control by government affects virtually every

business in the country in one way or another—usually in many different ways.

This may well have been the fundamental explanation for Watergate—that is, for the aspect of Watergate represented by corporations' illegal contributions to former President Nixon. As one corporate executive put it at one of the Conference Board conferences that a political scientist, David Vogel, and I covered and have described and analyzed in a forthcoming book, *Ethics and Profits*, "When it came down to Nixon or McGovern, the outcome really meant a lot. This was the fundamental reason for illegal political actions."

Yet, remarkably enough, at least from the perspective of most critics of business, businessmen do not see themselves in control of the political process. Quite to the contrary, just as much of the public sees powerful business corporations dominating the rest of the society and the governmental process, businessmen see just the reverse—they believe they themselves are dominated by other forces in the society—populist politicians and their supporters, government bureaucrats, labor unions, farm groups, citizen groups, the press and the electronic media.

A great many reporters, scholars and critics of business have assumed and asserted for a long time that business and government enjoy a very close rapport; but our observation has been that, whatever that rapport may or may not have been in the past, and it has doubtless had its ups and downs, business today is extremely suspicious of and hostile toward government. And, as business distrusts government, government distrusts business. In each case, this distrust is based largely on misconceptions regarding the roles and performance of the other party. This is a bad situation because many of the problems that our society will have to face require increased cooperation between government and business; this will be essential if the nation is to solve persistent problems of economic growth and stability, to end high unemployment and inflation, to reduce the social and international tensions that result from economic inequality and poverty, to check and reverse urban decay, to avert threats to the natural and social environment of an expanding industrial system, to accomplish the rebuilding of a healthy and decent social order in which business institutions, as well as those of government, can regain public respect.

One of the most striking findings of our study, we think, was the distrust that business feels not only toward government but toward *the democratic process itself*. This is a latent cause of much that has gone wrong in business's public standing.

Strikingly, businessmen today seem remarkably pessimistic about the future of the capitalist system. The only group that is even more convinced that we are witnessing the twilight of capitalism are the Marxists. The loss of faith by businessmen in the compatibility of capitalism and democracy could be a self-fulfilling prophesy. There needs to be an end to the kind of cynicism that leads to self-interested attempts to manipulate politicians and the public.

There is obviously no simple formula for how business can regain public respect and understanding, since what is involved are all the things that individual executives and corporations do in their relations with government and the public, both at home and abroad, as well as their internal conduct of their business affairs. What corporate executives need to accept is that they have *two* major roles to play: One directing and managing the affairs of their companies, the other in recognizing and responding intelligently to the expectations and needs of the broad society. If they neglect the second role, or despise it, they will get themselves and their organizations into deep trouble and deeper public disrepute, as some have already done. This is the lesson to be learned from the seemingly hard-headed, hard-nosed, narrowly profit-oriented behavior of some corporate executives in recent years; they seriously hurt themselves, their companies, business in general and their country as well as other friendly nations by neglecting or misconceiving their public role.

The resulting extremely adverse public reac-

tion is moving some leaders of business thinking toward acceptance of a new concept of the role of business in society—a concept that is far broader than the earlier ideology which held that the sole aim of the corporation is to produce a profit. Profit-making is obviously crucial to business survival and growth but it can no longer be celebrated as a sufficient objective by businessmen—not if they expect to regain public confidence and avoid highly constraining public regulation, control or even expropriation.

We have called the new creed that is emerging "the consent doctrine"—the recognition of the public's participation in shaping business policy and business actions. Businessmen must recognize that they play their role, exercise their considerable power, subject to the consent of the public. The ability of corporate executives to exercise their considerable powers effectively depends on their obtaining and holding this public consent.

This consent doctrine does not in my view imply a merely accommodating or passive role for business. Rather, business should seek to contribute actively, by its own performance, its policy advice, and its cooperation with other groups, to the solution of the grave problems that trouble society today and will affect it tomorrow. Businessmen must learn to look ahead and help government do what the individual corporation cannot do: tackle the broad, long-range problems that lie beyond the reach and grasp of the individual firm. In contributing to the broad social welfare, business will regain respect for itself, and for other participants in the democratic process.

Can business actually do it? Radical critics of capitalism insist that it cannot—that business institutions are inherently so narrowly self-interested that, in conducting the normal quest for profit, they must undermine or corrupt the democratic process. It is up to business leaders to prove that such a thesis is wrong—if they hope to preserve their freedom

It may in fact be true that, in many instances and in the short run, corruption pays; but sooner or later it will cost heavily. Over time, it will lead to the destruction not only of individual firms but of capitalism itself as government takes control. The survival of private business institutions and the values of independence and liberty that businessmen cherish thus depends not just on the ability of business to earn sufficient profits, which it does, but also on a broader and deeper conception by business of the public good. To command significant public support, a conservative ideology must provide a better defense of limited government than that of preserving the economic freedom, privileges, prerogatives, wealth and power of corporations and their managers or owners

At any rate, that is my own prediction and prescription, colored beyond doubt by my own ethical presuppositions.

Professional Standards for the Performance of the Government Economist

By JOHN B. HENDERSON*

No employee of an organization is entirely "his own man"—he is a part of a system, whether it be an academic, commercial or governmental system. There are different constraints, and one of the principal ethical questions is the individual's perception of, and response to, these constraints. What distinguishes the governmental or the business economist from the greater part of academic economists is that he occupies a staff role. Thus the fortunate few at the best of academic institutions are more nearly "their own men," in that professional standards evaluated by peers determine the norm of performance, with relatively little bureaucratic intervention in that judgment.

The position of an economist in the federal government, however, may differ only in degree from that of the academic economist. I, for one, would certainly not have consented to a role that was so defined and circumscribed as to make government economists mere creatures of the established order of things. But there are real constraints, and they depend, to an important extent, on the nature of the position that the economist holds. Equally, there are real professional obligations. It is my contention that government economists perform their most effective professional work when they speak plainly and do not trim their views to suit the presumed wishes of those who consult them; in other words, when they deliver their best analysis of reasonable policy options in fair and comprehensible terms.

This is easier said than done. There is a great diversity of professional positions in the federal government. In some cases, the professional commitment of economists is strong and is expected to be so; in others, it is not of paramount importance even if it is never unimportant. Some positions are relatively secure, and others are dependent upon the retention of political favor. Some organizations encourage freedom of expression when the doors are closed, and others count upon a rather stiff orthodoxy. At any rate, the penalties visited upon the nonconformist are not at all uniform.

The environment in which the federal government economist works is always affected in some way by the political process. Hence there is little room in the federal government for the professional economist whose style is utopian or dogmatic. The ethical question of how to dissociate one's personal experience and preferences from one's analysis is not answerable in an absolute way. But the general injunction that it is best to be candid and fair—and simple too, in order to be understood—seems to me to hold true for all federal government economists and to involve, occasionally, only the kind of risks that a professional should face and not dodge. It should be said, however, that for those economists who arrive in the federal government from positions in private business, candor and fairness ought to call for their making a very serious effort to avoid the continuing promotion of a special interest. The public payroll deserves no less service.

Let me then sketch some of the problems of judgment and conduct that are likely to face different groups of economists in the Federal Civil Service.

The first group, relatively few in number, consists of those who hold political appointments in the executive branch. It includes not only distinguished and visible officeholders, such as the members of the President's Council of Economic Advisers, but a diverse group of people, not only economists of course, who are

*Congressional Research Service, Library of Congress

appointed under Schedule C. These people neither go through the regular process of civil service recruitment nor enjoy any right of tenure. They are therefore dependent upon the Administration, and the expectation rightly exists that their views, on economic and other subjects, should be politically compatible with those of the Administration.

This represents one clear case of economics not being, so to speak, in a separate box. It is necessary, of course, that these should be people of technical and analytical ability, capable of working under acute pressures and of meeting severe time deadlines. But there are several ethical considerations that affect them. For example, how far should they go beyond the range of decently objective analysis? To what extent should they be not merely drafters of position papers, not merely cool advisers, but active advocates or even propagandists of their views within the Administration?

I admit that my opinion on these matters is a conservative one, with a small *c*. It is that it is prudent, and ethical, for the adviser to stay modest, not to underestimate the uncertainties, and to commend his analysis rather than to try to command its acceptance. Self-confidence on the part of the analyst is all very well, but enthusiasm is all too often unprofessional. I am not advocating a constant skepticism, but I judge that on past evidence those advisers have been most effective who have shaped their standards by those of the profession at large, who have adhered most strongly to the general public interest as the professional aim, who have taken the long view in an environment where the short term frequently offers the greatest political appeal and who resist the temptation, as professional economists, of becoming instead amateur politicians. There is, of course, a directly ethical answer to those who are tempted to become politicians. That is that they should try, and see how far they are able to go.

There is another, extremely difficult, question facing the political appointee. Someone in authority will listen to his analysis. It will not always be accepted, but, if it is respected, that in itself is enough. If the opposition on other grounds, probably political, is overwhelming, then the economist should accept it. The question arises, at what point to cry halt. When should the professional economist insist upon his advocacy at the risk of his being fired? When should he decide to quit?

There is no general rule that can be declared on this. But I find it regrettable that resignation from public office is so little used these days—and not only in the economics profession—as a way of making a protest. Yet I can understand why it is so.

One reason is that the economics profession has been so little interested in making this definitive action, of resignation, an easy option. The nonacademic economist, on resignation, has no easy recourse to joining a partnership. But academic economists, spending their sabbaticals in office, can at least regard their possible premature return to the university as an escape hatch. Even that is not often used, such is the charm of office.

The second reason for not resigning is, in my view, less creditable. It is the self-serving myth that to stay in office and to work within the systems is a better contribution than to quit in protest. On small matters this may be a valid claim, but on larger issues I find it difficult to accept. For example, the massive turnaround of domestic economic policies in August 1971 towards wage-price controls was accompanied by little or no change in the group of economic advisers. In some cases, the economists' opposition to controls of any kind had been so assertively stated that it was doubtful that they could continue to be sympathetic advisers for their effective administration. That seems to me to be a case where public confidence in the administration of a changed policy called for a changing of the guardians of that policy.

Let me come now to a different, and also relatively small, group of professional economists employed on the payroll of the United States Senate or the House of Representatives. They too are not appointed in the ordinary process of civil service recruitment, and they owe their

allegiance to a Senator, to a Congressman, to a Committee Chairman or at the least to an administrative assistant or staff director. This is a perfectly acceptable political process, compatible with the uncertainty and terminability of elective office. But it tends to discount professional excellence and to demand instead a steady obedience to political prescriptions. This, I believe, is why, until recently, the quality of economic advice on Capitol Hill has in only a few places been worthy of professional recognition.

It is the exceptions that provide for me the greatest encouragement. In the not so distant past Senator Douglas, as Chairman of the Joint Economic Committee, initiated in 1959 a study of Employment Growth and Price Levels that was marked by its wide view of the issues, its long time range and its thoroughly professional quality The Joint Economic Committee neither had nor has legislative oversight. But that study contributed, in a significant degree, to the acceptance by the Congress of a shift towards compensatory fiscal policy in the early 1960's, in much the same way as the professional advice of the Council of Economic Advisers had promoted a wage-price-productivity rule that contributed to a stable recovery.

Skilled analysis, even when it is broadly accepted within the profession, does not have a guarantee of immediate approval. It is in this context that the federal government economist looks for professional support to the relatively small group of academic economists who, as consultants and expert witnesses, regularly reinforce the process of professional persuasion. For the professional within the federal government is most often competing for attention with those who are serving political interests. The Congressional economist, in particular, is never very far away from representations of all kinds by constituent interests, more or less well organized. If they are well organized, they are called lobbies, seeking the promotion of special concerns, trying to justify preferential treatment, and claiming, in the area of taxes, the kind of consideration that opponents would call loopholes. How does a professional economist cope with such a situation? There is no use in merely deploring the existence of lobbies. They can be countered and opposed, but they cannot be wished away. The economists of the federal government have to recognize that lobbying is part of the political process and have to place their confidence in the thought that the sober and fair evaluation of the issues will eventually prevail and that the general public will come to recognize where the general interest lies. Ultimately, it is an act of faith that economists' contributions to government will be vindicated and that economic wrongs will be set to right.

This means that professional economists, in both the executive and legislative branches, tend to be the natural, but not always effective, adversaries of the special interest lobbies. I doubt whether economists in the academic field who do not have regular contact with the federal government understand how emphatically the public interest calls for the defense of unorganized but vitally interested parties such as the consumers, or to what extent the great majority of professional economists in the federal government and the few protagonists of the general interest like Allen Ferguson stand as their defenders. Not all the "meetings of merchants," in Adam Smith's phrase, are reprehensible or even preventable, but it is surely true that they constitute in many cases a "conspiracy against the public." If the lobbies are successful, the legislative impact is always plain after the event, for example, in the enactment of tax favors or the rejection of federal no-fault insurance.

On the other hand, the legislative branch in recent years has increased the number of economists who owe no obligation of party allegiance. The economists of the General Accounting Office and of the Congressional Research Service in the Library of Congress are answerable to the entire legislature and not merely to the majority or minority. This confers a very great potential advantage in terms of professional quality of performance. And the relatively greater job security of the economists in

these agencies, as compared with their counterparts on the Senate or House payrolls, ought to be associated with a more highly professional product.

Yet perhaps the most valuable development that has occurred in the legislative branch in the past few years has been in the professionalization of the Congressional budget process. While the Congressional Budget Office (*CBO*) is primarily and directly answerable to the two Budget Committees which are, of course, political bodies organized by the majority party, the professional quality and evenhandedness of the *CBO* product is such as to meet fully the ethical demands of the profession and to provide, as it should, the starting point of a political discussion. Nothing in the way of policy decision is prejudged by the analyses of the agency. But the opportunity thus offered to the Congress to address the fiscal situation as a whole is something that I regard as immensely valuable for the development of legislative initiative, and, for that reason, as making a contribution of the highest professional quality.

So far, I have not mentioned the great majority of economists in the federal government. These are the career civil servants in the executive departments, and they comprise a great variety of people who call themselves economists These people have been appointed by the competitive process and they are thereafter assured, if they so desire it, a security of position that is substantially better than in most of the private sector.

What are their problems, and what are their ethical obligations? First, it must be accepted that they are the staff of an established order and that whatever initiatives they invoke—and they are limited—must be subject to political constraints. In the executive departments, which can be regarded as representative of broad constituent interests—commerce, labor, transportation, agriculture and the like—the economists of each department are virtually required to shape their attitudes according to departmental style—a situation that is very closely comparable to that of corporate or bank economists. If that is so, and the security of their position depends upon their adherence to a sectoral orthodoxy, then there seems to be, in this, the making of an entrenched bureaucracy Is it really so? I cannot deny it categorically. And yet I believe that the professional quality of the economists in the executive branch is such that they are more vigilant in the public interest and more sensible of the need for farsightedness than some of the more cynical commentators would allege.

I am aware that this estimate of the performance of the average economist of the federal civil service is more complimentary than the commonly accepted image. But I ask you to accept the thought that ethical, professional and meritorious performance by the federal government employee is not measured by intellectual brilliance nor by zeal, but by clarity, moderation and balance. A civil servant will be effective to the extent that his views are accepted as having professional authority, and that will almost never be achieved by dazzling the people who are being advised Furthermore, there is always the doubt about where the public interest truly lies. I am impressed by the validity of the statement of the late Kermit Gordon quoted by Leslie Gelb in his affectionate obituary: "Men possessed of strong analytical powers—men of goodwill, disinterested men—will often define differently the public interest in a particular problem."[1] I believe that that situation occurs, more often than most people recognize, within the executive branch.

And so I think that, mercifully, the focus of the ethical issue for the career civil servant has shifted away from the question that Leonard Silk addressed so eloquently in his paper, "Truth vs. Partisan Political Purpose," before the American Economic Association almost five years ago. At that time, it was not merely the entitlement of professional economists within the federal government to provide objective analyses that was under assault, but the veracity of official statistics. In these circumstances, a vigilant press is essential, for the civil servant himself cannot provide his own defense. And it will always be in the interest of an open society to have the performance of politicians and professionals alike subjected to re-

[1]*New York Times*, June 23, 1976

view and commentary by the press. In the past two years, by contrast with that earlier time, there have been relatively few occasions when the career civil servant has had his evaluations of statistical evidence second-guessed by a variety of political interpreters. But the line between analysis and political interpretation is not easily drawn, even if there is no longer such an acute dilemma between telling the statistical story as it is, and offending the political powers in so doing.

On the other hand, I believe that the federal government economist's attitude towards the development of programs that involve the spending of money and the employment of civil servants poses an ethical issue that is just as acute now as it has been in the past.

The progress of a program through the bureaucracy is almost predictable in general terms. A new need is perceived. The means of achieving it are debated, and one alternative is chosen, eventually with political assent. This is the expression, however imperfect, of the public interest, and the staff people, including the economists, are expected to go along. Most of them do, willingly, as contributors to the original concept.

But what happens when something is amiss? When the economist fundamentally disagrees with the aim of the program? When it appears that the primary result is that the bureaucratic managers are justifying and defending their existing positions, instead of devoting resources to the purpose of the program? It is obvious that some provision has to be made for administrative expenses and for the recording of performance. But at what stage does that expense come to be devoted to the defense of the existing order? Does the bureaucracy perpetuate its existence in this way? And what cost does this impose in terms of professional performance?

I wish I could give an unqualified denial that bureaucracy is wasteful. I cannot. But, whether he is aware of it or not, the permanent civil servant usually becomes involved in the spirit of the organization, much as the corporate economist, the bank economist, or any other committed employee.

Yet it is impossible to ask that a career civil servant should react to bureaucratic delays or inefficiencies or wastefulness or errors by quitting the service. What then is proper? If the dissent within the organization is apparently manageable, let him (or her) take it as far as advocacy, behind closed doors, can manage. If it is effective, so be it. If it is not, then let him (or her) find, quietly, and probably slowly, another place and another program.

What this leads to is the thought that the diversity of the federal government is its saving grace. The sensitive and moderate economist, who does not aspire to executive power, is not, in most cases, going to find that his situation is intolerable. He has to regard himself as being in the role of an expert adviser.

It is, however, usually possible for a staff economist to stay within the bounds of economic analysis and interpretation, and yet exercise the power of persuasion in the choice of alternatives. If he refrains from claiming the role of policy maker, I see no reason for his not urging the adoption of ideas that he approves.

Nevertheless all economists in professional positions in the federal government should accept the view that representative government assigns responsibility for decision-making to elected officials and that unelected advisers are subordinates. Hence the economist's professional contribution can be well made even if it is not accepted, for the economics profession has no natural monopoly on the offering of advice

On the contrary, the rule of public morality that seems particularly applicable to the United States is that all opinions should have a hearing and that professionals should have no special privilege. Hence the key to professional effectiveness is the ability to explain matters more persuasively For the federal government economist, it means that ethical performance and plain speaking are closely related to one another.

REFERENCES

Leonard Silk, "Truth vs. Partisan Political Purpose," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 376–78.

Environment, Health, and Economics—The Case of Cancer

By ALLEN V. KNEESE AND WILLIAM D. SCHULZE*

It is becoming recognized in the United States that nutrition in the broadest sense, that is, all substances that are ingested, is a very important factor in health, perhaps the most important over which man has any control. To some extent this insight results from observation of relations between geochemistry and health.[1] For example, the relationship between iodine deficiency and endemic goiter has been recognized for many years. Similarly the relationships between iron deficiency and anemia and between flouride deficiency and dental caries are well known. Others include zinc deficiency and dwarfism, and lithium deficiency and mental disorders. As such relationships become better understood, large opportunities to improve health through preventive measures become available. It is most unfortunate that medical research has devoted itself almost entirely to "cures" rather than understanding the etiology of disease.

Even more urgent perhaps than understanding how natural geochemistry influences health and disease is finding out how human interventions in the chemistry of the environment are related to disease. There are many such effects possible and a number have occurred on a relatively large scale. Dramatic instances of heavy metal poisoning have happened, most notably in Japan. A number of combustion products discharged to the atmosphere from stationary and mobile sources are implicated in lung disease. But the suspected relationships between environmental contamination and cancer are currently receiving the most attention because in many ways that is our most important disease.

Cancer incidence has been rising at an average rate of approximately 1 percent a year for some time. Heart disease is still the largest killer, accounting for about 40 percent of total deaths in 1975 according to the National Center for Health Statistics. But the death rate from heart disease went down while cancer deaths, about 20 percent of the total, went up more than 2 percent in 1975. Contrary to the impression which might be gained from some of the more dramatic stories in the press, the long term rise has been dominated by only one form of this disease, lung cancer.

Since 1930 lung cancer has been rising steadily for both men and women but much more rapidly for the former. The recent upturn in lung cancer rates for women reflects the delayed (relative to men) popularity of cigarette smoking among women. Thus, the increase in female smoking in the 1930's and 1940's may shortly cause a dramatic increase in female lung cancer rates. Stomach cancer has been trending pretty steadily downward, although quite recently cancers of the digestive tract have shown an increase and the others have been constant. The increase in lung cancer appears, as mentioned, to be strongly related to smoking. Lest one feel complacent about the situation, however (all we have to do is stop smoking and anyway smoking is voluntary, one might say), one should recall three sets of facts. First, the cancer rate is high in the United States, accounting for a large portion of mortality, and death from cancer is often especially painful. Second,

*Professor of Economics and Associate Professor of Economics, respectively, The University of New Mexico.

[1] For a review of the evidence, see *Geochemistry and the Environment, I*, National Academy of Sciences 1974.

cancer is now widely recognized as being, at least mostly, perhaps entirely, environmentally induced with a latency period between 15 and 40 years. Third, because of this long lag we have not yet seen the effects, whatever they may be, of the massive introduction of chemicals into the environment and into work places, which followed World War II.

On the last point the Council on Environmental Quality concluded, "About two million chemical compounds are known, and each year thousands more are discovered by the *U.S.* chemical industry and hundreds are introduced commercially. We know very little about the possible health consequences of these new compounds. Many are not toxic, but the sheer number of chemical compounds, the diversity of their use, and the adverse effects already encountered from some make it increasingly probable that chemical contaminants in our environment have become a significant determinant of human health and life expectancy."[2]

The evidence that cancer is environmentally induced is at this point largely circumstantial. The regional pattern of cancer mortality is suggestive. For total cancer in males one sees in the cancer atlas that in general high rates are found in the larger cities with particularly heavy concentrations in the Northeast and the lower Mississippi Valley. If the rates of cancer everywhere could be reduced to the lowest ones actually experienced anywhere, cancer mortality would drop by about 90 percent. We badly need to understand the mechanisms involved and the costs and benefits of preventative approaches to the problem. We will discuss one study which tries to provide such information a little later.

Since the right basic approach to cancer appears to be prevention, although we also need to learn how to treat it better, the heavy orientation of medical research toward "cures" is misplaced. Furthermore, the massive research in recent years has not even resulted in much improvement in survival rates. The greatest improvement in these rates occurred in the 1940's and 1950's probably because of advances in detection, treatment, and surgical technique. Since the 1950's three-year survival rates have shown little change. Our national cancer research program badly needs reorientation. A large proportion of the money spent on it so far has been wasted.

I. Some Implications for Economics

Having criticized the orientation of medical research, we must hasten to add that research results in the area of health economics in general and cancer in particular are also rather meager. This is mostly because of lack of attention to these problems by economists. The profession never seems to tire of investigating the term structure of interest rates while being oblivious to enormous social problems to which the economist's skills could be constructively applied.

Nevertheless a good deal has been accomplished in environmental economics and the field is active.[3] But the study of cancer presents particularly difficult problems for economic analysis. The long latency period of the disease means that it is extremely difficult to link cause and effect. Further, the combination of this delayed effect and the fact that the probability of contracting cancer when exposed to a carcinogenic substance appears to be (probably linearly) cumulative over time means that any optimization approach to the problem must be dynamic.

The quantitative economic models which have been developed to analyze environmental problems have in general been either static, steady state, types, or if time dependent have been deterministic and incapable of handling cumulative phenomena. There has been at least one application of an optimum control model to cancer which recognizes both

[2]See *Environmental Quality, The Sixth Annual Report of the Council on Environmental Quality*, USGPO, Dec. 1975, p. vii.

[3]The high quality articles in the *Journal of Environmental Economics and Management* reflect this. Also there is a steady stream of books being published in this field.

the latency and cumulative risk aspects, but such models have, so far, not lent themselves well to quantification. The theoretical conclusion of this model is that if the carcinogen has a positive value to those ingesting it, it is optimal to control exposure to it more at younger ages and less at older ages. This is an interesting but not very surprising conclusion (Reza Pazand 1976). However, there also are some more empirical studies which are needed (and some of which are underway). We will consider these under the categories of cost studies and damage studies.

We know very little about the costs of controlling carcinogens in the environment. In the United States the main element in the problem, aside from smoking, appears to be the explosion in the manufacture and use of chemicals. Very few of these have ever been tested for carcinogenesis. Each year there may be about 6,000 new substances to which humans are exposed. There are already about 1,500 which are suspected to be carcinogenic. One reason that so few substances have been tested is that an animal test worth anything at all costs about $150,000. Another is that the burden of proof has been on the government to show that a given substance is a cancer agent rather than on the marketer to show that it is not. It appears that the cost of screening is likely to go down drastically because of the availability of a new testing procedure. This opens a new opportunity for economic research and policy formation which we will suggest shortly.

The new procedure is called the Ames test and uses a particular single cell bacterium. One billion bacteria are used in the test, which costs only about $500 per substance. The test actually identifies mutagenesis, which is thought to be related to cancer.[4] In one experiment 174 substances thought on other grounds to be cancer inducing in humans were tested and the Ames test identified 156 as carcinogens. Thus 90 percent of the suspected carcinogens were identified as such by the test. Of 46 common biochemicals believed not to be carcinogenic none was identified as such by the Ames test. In addition, the test can identify carcinogens on a very broad scale. Thus a chemical which might induce cancer in even several tens of thousands of Americans could never be identified by any feasible animal tests. The colonies of test animals would simply have to be too large. Since the Ames test uses a billion subjects it can identify such a chemical.

An interesting and useful piece of biological-economic research would be to take all the chemicals introduced in a given year that might possibly be regarded as carcinogens and screen them using the Ames test. Even if all 6,000 were tested this would only cost about $3,000,000, a tiny fraction of what the researchers at the National Institute of Health spend chasing apparently nonexisting viruses. Probably only a fraction of the 6,000 would need to be tested. The subset which is found to be under suspicion should then be analyzed from an economic standpoint. The question would be how many of them are really of major economic value which justify subjecting the society to some risk. Our hypothesis is that nearly all of them would be found to be of quite marginal value. It appears probable that the chemical companies are engaging vigorously in product differentiation. If this biological-economic experiment proved successful a new approach to control of carcinogenic chemicals could be built on it.

Under this approach any one who proposed to market a new chemical would be required to subject it to the Ames test. If this test indicated carcinogenicity at any level the potential marketer would be required to prepare an economic impact statement on the material. This statement would have to indicate what functions the chemical performs, how valuable these functions are, what substitutes exist for it and data on the performance and costs of these substitutes. These statements should accompany the application for marketing the chemicals. It is economically illogical to treat, in the regulatory process, each chemical as though it had equal value to the society, as is done now. In

[4]A discussion of this and related tests can be found in *Science*, June 18, 1976, p. 1215.

addition the burden of proof should clearly be on the marketer to bear any costs of damage which the chemical might ultimately cause. These policy changes might greatly reduce the number of chemicals brought on the market each year. But if our hypothesis is correct the social cost of this protection would be small.

But this is not to say that protecting against carcinogens will always be cheap. Fairly ordinary, well established, and widely distributed chemicals have been found to be carcinogens—this includes polyvinylchlorides and vinyl chloride to name a few.

In this connection nitrogenous compounds are a particularly interesting example. Nitrosamines, which are readily formed in the body from nitrites and secondary amines, are under severe suspicion as cancer agents. Many nitrosamines are powerful carcinogens to animals even at low concentrations. Nitrates and nitrites are used as additives and preservative agents in the food industry. They prevent bacteria growth and preserve the color of meats and meat products. One reason they are under suspicion is because of observed high rates of stomach cancer in countries where much preserved meat is eaten. Nitrosamines have been shown to cause cancer in many other parts of the body too.

There are also environmental sources of nitrogenous compounds. In some locations the level of nitrates in drinking water is high, and nitrogen oxides are emitted in very large amounts from the stacks of power plants and automobile exhaust pipes. The mechanisms whereby nitrogen compounds are converted into nitrosamines is not very well understood but precursors such as nitrates have been converted into nitrosamines in the digestive tracts of test animals.

In connection with a study of nitrogen compounds being conducted by the National Research Council, members of the Resources Economics Group at the University of New Mexico have conducted a study of possible damages resulting from cancer induced by the ingestion of nitrogen compounds and other environmental factors (Pezand, forthcoming). To determine such damages, two relationships must be known. First, one must understand the relationship between the occurrence of nitrogenous and other carcinogenic compounds in the environment and the occurrence of cancer. Secondly, one must know something about the value of the lives lost due to the induced cancers.

A little about the etiology of nitrosamine induced cancer was said earlier. Econometric techniques were used to test hypotheses concerning the occurrence of nitrogenous and other compounds in the environment and cancer. Notable features of this study are that it included numerous suspected explanatory variables simultaneously in an effort both to achieve a fully specified equation and to include some measure of the total body burden. It also included lagged explanatory variables to account for the long latency period of cancer, and it accounts for the effect of population mobility on observed cancer mortality rates for cities used in the analysis. As usual, data problems were severe. Constructing the data set was the most trying part of the research, especially obtaining the historical data for the lagged variables reflecting exposure to nitrogenous compounds. It is of great interest to note, however, that some of these variables were not significant when introduced into the estimating equation at their current values but were highly significant when lagged sixteen years to reflect the latency period of cancer (sixteen years was the longest lag possible with the data set).

The dependent variables, expressed in terms of deaths per 1,000 of population, are 1972 mortality rates for cancers of the: 1) digestive system, 2) respiratory system, 3) breast, 4) genital organs, and 5) urinary system; and mortality rates for: 6) leukemia, 7) all other cancers, 8) all cancers, and 9) total mortality. The independent variables used fall into four categories: socioeconomic, air quality, water quality, and life style. It was possible to construct the full data set for 60 cities.

Complete regression results are presented in Table 1, which summarizes results of the study noted above and shows the significance of variables which were correlated with the occurrence of cancer. The beef consumption, pork

TABLE 1—ENVIRONMENTAL FACTORS IN CANCER MORTALITY
(mortality in deaths per thousand)

Independent Variables		Linear Regression Equations[a] Estimated Coefficients with t-Values in Parentheses[b]								
Name	Mean Units	Mortality	All Cancers	Digestive	Respiratory	Breast	Genital	Urinary	Leukemia	All Other Cancers
% Change in Population (1960–1970)	—	− 048 (−2 02)	− 0033 (−1 16)	− 0027 (−3 13)			− 0012 (−2 46)			
Median Age of Population (1970)	Year	52 (3 38)	104 (7 25)	024 (4 20)	025 (6 65)	01 (4 35)	01 (3 06)		.0059 (4 95)	026 (5 49)
% Nonwhite of 1970 Population	—				000086 (1 37)	000054 (1 27)			− 00002 (−1 08)	− 00008 (−1 09)
1969 Mean Family Income	Dollars		− 000003 (−1 2)		− 0000012 (−1 8)					− 000001 (−1 76)
1959 Per Capita Beef Consumption	1 4 Lb/Person/ Week	−2 86 (−1 34)						039 (1 79)	034 (1 66)	
1959 Per Capita Pork Consumption	1 09 Lb/Person/ Week		68 (2 36)		18 (1 89)	108 (1 83)	204 (2 99)		079 (2 61)	22 (2 06)
1956 Per Capita Cigarette Consumption	110 0 Packs/ Person/Year			002 (2 85)	00074 (1 33)	00085 (2 31)				
NO_2 (in the air)	1 15 Micrograms/m^3								− 18 (2 05)	
SO_2 (in the air)	43 04 Micrograms/m^3								− 00008 (−1 73)	0003 (1 90)
Ammonium (in the air)	09 PPM		07 (2 89)			02 (3 01)				032 (3 42)
Suspended Particulate	114 8 Micrograms/m^3			00062 (1 48)	00046 (1 63)		00035 (1 45)			
Sulfate (air)	10 79 Micrograms/m^3					− 0057 (−3 03)		001 (1 54)		
Beta Radioactivity	$26pC_i/m^3$	9 25 (2 14)							057 (1 73)	
UV Radiation	18 79 Microwatts/cm^2	052 (2 26)								.001 (1 49)
Nitrate in Drinking Water	2 45 PPM	15 (1 67)				0023 (1 56)				
Constant		−3 11	−1 82	436	− 599	− 32	− 31	019	22	− 64
R^2		38	71	63	58	41	45	09	44	63
SSR		367 7	2 902	5464	2439	095	1733	044	019	2619
DF		53	54	55	53	52	56	57	46	46

[a]Independent variables with t-values below 1 0 excluded from final estimated forms
[b]Significance level: 85%: $1.04 < t < 1.28$; 90%: $1.28 \leq t \leq 1.68$; 95%: $1.68 \leq t \leq 2.33$; 99%: $2.33 \leq t$ (one tailed t test)

consumption, and smoking variables were lagged. Unfortunately data did not permit lagging the other environmental variables. The usual caution must be stated that these are merely statistical relationships and cannot conclusively demonstrate cause and effect. Still, the variables were carefully chosen and a number of strong partial correlations are evident. Beef consumption, pork consumption, cigarette smoking, and ammonium in the atmosphere exhibit strong positive correlations with a number of types of cancer, and all involve the ingestion of either nitrosamines or their precursors. Several carcinogenic nitrosamines have been found in tobacco smoke. Meat in general is high in nitrates, and pork especially so because of the way it is processed.[5] Pork consumption shows an especially strong relationship with several types of cancer. Both ultraviolet radiation and beta radioactivity are also shown to be significantly correlated with cancer mortality.

How does one convert these findings into damages, assuming that they do reflect real

[5]The picture is clouded somewhat by the use of other additives in cattle feeds and certain meat products.

relationships? One necessary ingredient is an estimate of the value of life. One measure that has often been used as reflecting the value of life is the present value of lost earnings. This measure can perhaps be regarded as an extreme lower bound of the correct value, but it is so far off the mark that it cannot be considered very useful. It accords no value to the lives of housewives or old people, whereas an employed person may regard the value of the life of a spouse or parent as being fully as high as his own. We want in fact to know either how much a person would have to be compensated in order to accept a small increase in the probability of untimely death or how much he would be willing to pay not to be subjected to it, depending on the situation.

In a recent paper Richard Thaler and S. Rosen have provided some evidence on an equilibrium market measure of risk, valid for small increases in risk. By using regression analysis and information on the riskiness of various jobs they found that jobs with an extra risk of mortality of .001 pay $260 more per year. Since this level of increased risk implies that one man extra out of a thousand would die each year in the riskier job and since each one of the thousand must be paid the extra $260 this implies that the total premium paid for the life lost is $260,000 (Thaler and Rosen, p. 36). If one takes this as a first approximation it is possible to combine this result with information from the regressions reported earlier to come to some preliminary estimates of damages associated with carcinogens in the environment. Before proceeding to do this, however, it should be noted that the Thaler and Rosen figures are probably seriously biased downward. First it is not unreasonable to suppose that persons who accept risky jobs are less risk averse than the general population. Second, observation suggests that people in general are much more willing to take risks voluntarily than to have them imposed externally, consistent with the concepts of equivalent and compensating variation. This difference as well as the variation of either measure with changes in the probability of death is excluded from the Thaler and Rosen study. Environmental carcinogens fall at least partly into the involuntary category. Finally cancer is frequently a particularly slow, painful and unpleasant way to die, and the Thaler-Rosen calculation takes account of neither these effects on the suffering individual nor the external effects on other parties.

But let us accept Thaler and Rosen's figure as an acceptable first approximation of a value of the risk of death for small changes in that risk. In the regression equations relating various kinds of possible sources of carcinogens, the coefficients of the variables multiplied by the mean values of those variables give the probability that a typical individual (in the urban population) will contract some form of cancer. This figure can be multiplied by the Thaler-Rosen estimate of the value of life to get a valuation of the risk to the individual. This in turn may be multiplied by the urban population (150 million) to get the total valuation of the risk A similar procedure can be followed with respect to the other variables. When these calculations are performed for each of the variables where nitrogenous or other environmental factors may be involved and where the level of significance is at least 95 percent, it is found, as shown in Table 2, that damages associated with beef consumption are about 4 billion dollars per year, with pork consumption over 30 billion, with cigarette smoking about 12 billion, and with ammonium about 2 billion. Total damages from all sources are about 53 billion dollars per year. The full study reports many additional calculations pertinent to particular cancers. But these examples serve to illustrate the technique and the results. They suggest that we are dealing with large magnitudes indeed.

II. Concluding Comments

Fortunately the importance of cancer prevention is getting increased attention in the research and public policy community. One element in such prevention, perhaps the chief one, is avoiding the insertion of carcinogens into the environment. In some cases this can be done at low cost. In other cases the cost may be very heavy.

It appears that new testing techniques will permit complete screening of new chemicals for carcinogenesis. It will probably be possible to

TABLE 2—SUSPECTED DAMAGES FROM ENVIRONMENTAL FACTORS WHEN SIGNIFICANCE OF RELATIONSHIP IS AT LEAST 95%

Environmental Factors	Cancer Mortality: Digestive	Respiratory	Breast	Genital	Urinary	Leukemia	All Other Cancers	(Total Cancer Mortality) ×	(Mean Value) =	(Death/1000 per Population)	Death for Urban Population	Suspected Damages ($10⁶)[1]
Beef Consumption					039	034		073	1 4	1022	15,330	3 986
Pork Consumption		18	108	204		079	22	791	1 09	8622	129 330	33 626
Cigarette	002		00085					0029	110	3190	47 850	12 441
SO_x							0003	0003	43 04	0129	1,935	503
Ammonium			02				032	052	1 15	0598	8,970	2 332
Radioactivity						057		059	26	0153	2 301	598
TOTAL												53 485

detect potential carcinogens where the probability of causing cancer is quite low. It may not be desirable (optimal) to reduce this risk to zero, as the Delaney Amendment requires. New substances should be subjected to an explicit economic analysis as well as a test for carcinogenic properties. To treat all new chemicals as though they were of equal value as is presently done is not economically rational.

As we become more aware of the complexity and subtlety of the processes leading to cancer, and of its environmental origins, it becomes necessary to devise new techniques and methodologies for understanding these phenomena Econometric techniques can make an important contribution, but the effectiveness of their application is severely limited by data problems. In view of the magnitudes of damages apparently resulting from cancer mortality, it would be socially worthwhile to expend large amounts of funds to improve the data base

Economic analysis can provide defensible, lower limit estimates of damages if the risk is understood. But work on health damages by economists is still in its early stages and is an important area for further work in the developing field of environmental economics.

REFERENCES

Reza Pazand, *Environmental Carcinogenesis: An Economic Analysis of Risk,* Ph D. Dissertation, University of New Mexico 1976.

———, **Allen V. Kneese and William D. Schulze,** "An Economic Analysis of Carcinogenesis. What Cost the Risk?" Working Paper Series, Program in Resource Economics, University of New Mexico, forthcoming.

R. Thaler and S. Rosen, "Value of Saving a Life: Evidence from the Labor Market," unpublished 1975

National Academy of Sciences, *Geochemistry and the Environment, I,* 1974.

Science, June 18, 1976.

U.S. Government Printing Office, *Environmental Quality, The Sixth Annual Report of the Council on Environmental Quality,* Dec 1974, p. vii.

Incidence of the Benefits and Costs of Environmental Programs

By ROBERT DORFMAN*

This paper is concerned with the distribution among segments of the population of the benefits and costs of programs for protecting the environment. The matter is important from several points of view, including political viability and simple equity as expressed, for instance in Knut Wicksell's principle of just taxation. The paper consists of two parts. In the first we present some indications of how the benefits and costs of the current program impinge on different segments of the population classified by income level. In the second part we urge that the distribution of benefits and costs be taken into account in selecting and designing programs, and offer a suggestion for doing this.

Table 1 is a reminder of the magnitudes involved, which are far from negligible. It presents Department of Commerce estimates of the cash flows incurred for pollution control and abatement in 1972, divided according to the broad sector on which the costs fall initially. It should be read with the understanding that several major components of the Federal pollution control program were just getting under way in that year.

TABLE 1—POLLUTION CONTROL EXPENDITURES
UNITED STATES, 1972
($ billion)

Sector	Current Account	Capital Account	Total
Households[a]	2.0	.9	3.0
Business	4.1	4.0	8.2
Government			
Federal	.9		.9
State and local	2.1		2.1
Government enterprises	1.1	3.5	4.6
Total	10.2	8.4	18.8

Note: The federal pollution abatement program was in only partial operation in 1972.
Source: Council on Environmental Quality, *Sixth Annual Report*, Dec. 1975, p. 526.
[a]Including increased cost of home operation.

The total outlays in 1972 amounted to about $19 billion. Nearly half consisted of costs imposed on business enterprises by various environmental protection regulations, more than two-thirds if government enterprises be included along with private businesses. The remaining third was divided about equally between costs imposed directly on households and expenditures of governments of all levels. It is anticipated that for the next ten years at least the level of expenditure will be maintained at about the level indicated for 1972.

All the costs, whatever their point of initial impact, are defrayed ultimately by households. The governmental expenditures are transmitted by increases in taxes and by reductions in governmental services of other sorts. Business expenditures are shifted to households primarily by increases in prices. Both of these processes for shifting the burden from the initial point of impact to the ultimate payer are very complicated. The requisite analyses have been made by the Public Interest Economics Center working under contract with the Environmental Protection Agency. Our interest now is not in the technique of estimating ultimate incidence, which is intricate and requires a number of questionable simplifying assumptions, but in the results, some of which are shown in Figure 1. This figure shows the cost to households of different income categories imposed by the pollution control programs operative in three years, 1972, the current year, and 1980. The cost is measured as a percentage of family income. A perfectly neutral sharing of the burden

*Harvard University.

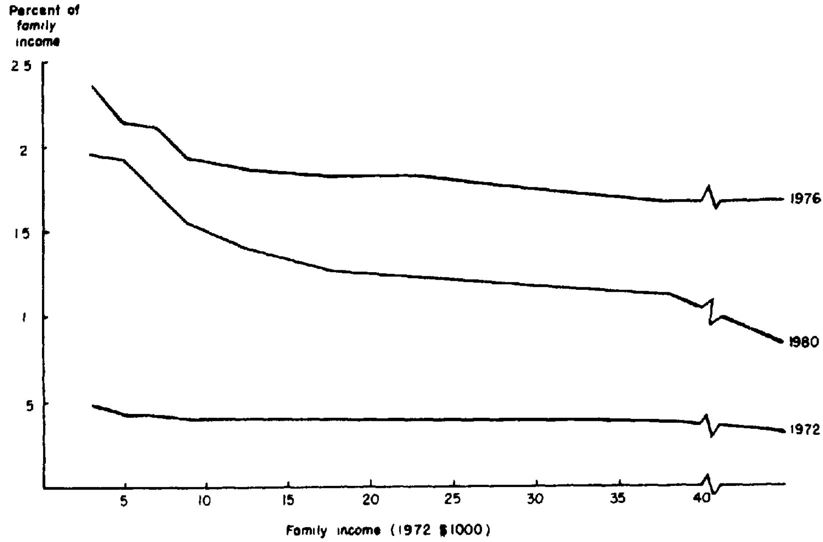


FIGURE 1 COST OF POLLUTION CONTROL PROGRAM, SELECTED YEARS, AS PERCENT OF FAMILY INCOME

Source Public Interest Economics Center

would be represented on the graph by a straight horizontal line. The sharing in 1972 was essentially neutral. In 1976, however, and especially in 1980 the distribution of the burden is distinctly regressive.

Figure 2 gives some insight into how this comes about. It shows, for 1976, the ultimate incidence of the costs borne initially by governments and industries and also the incidence of the costs of the mobile sources or automobile program, which is the major component of the costs borne by households from the outset. The costs of the mobile sources program are seen to be strongly regressive, and the costs borne initially by industries are almost as regressive. It is unfortunate that the data do not divide the costs attributed to government expenditures according to whether those expenditures are made by the Federal or by state and local governments. The total shown has a progressive impact. If the division were made it would show even more progressiveness for Federal expenditures and either neutrality or some regressiveness for the expenditures of state and local governments

The benefits of pollution control programs are even more difficult to estimate either in toto or as distributed among segments of the population. We do not have to review here the diversity of the benefits of those programs, the difficulties of estimating their magnitudes, or the obscurity of their values to different segments of the population. All available estimates are open to serious question. For our purposes the most appropriate and inclusive estimates appear to be ones derivable from some surveys made by the Gallup Organization on behalf of the National Wildlife Federation, which has provided the survey results to us. In 1969 the National Wildlife Federation survey of public opinion about environmental matters included the following question: "Would you be willing to

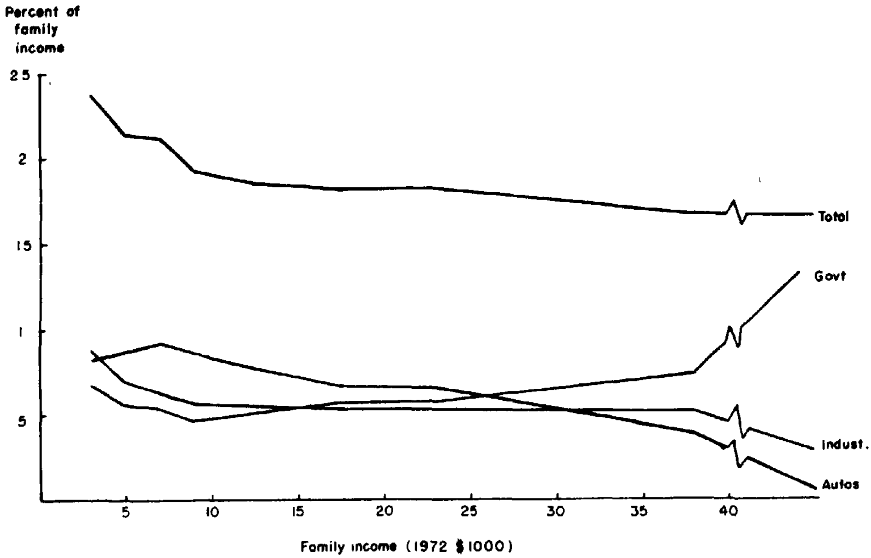


FIGURE 2 COMPONENTS OF COST OF POLLUTION CONTROL PROGRAM, 1976, AS PERCENT OF FAMILY INCOME

Source Public Interest Economics Center

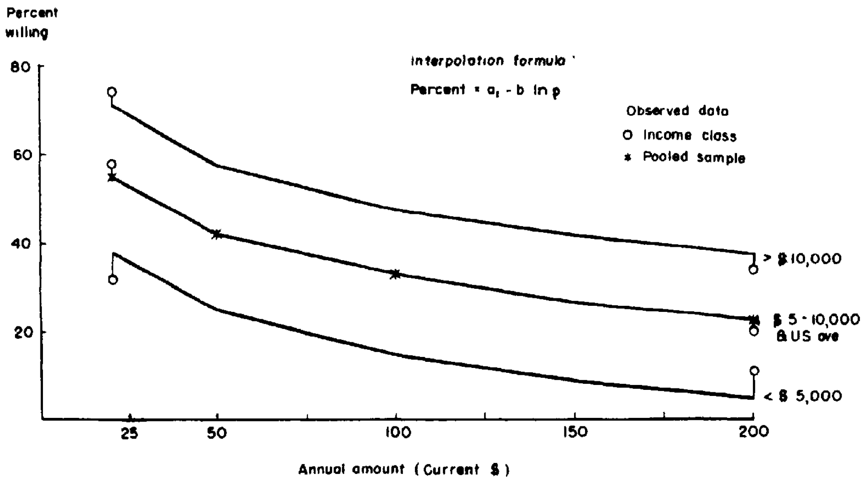


FIGURE 3 ASSERTED WILLINGNESS TO PAY FOR ENVIRONMENTAL CLEANUP, 1969 SURVEY

Source: National Wildlife Federation and interpolated data

accept a $X per year increase in your family's total expenses for the cleanup of the natural environment?" The question was asked for $X =$ $20, $50, $100, and $200. The tabulations available show the percent willing at all four levels for the survey population as a whole, and the percent willing to pay $20 and $200 for three income ranges.

Figure 3 shows the results. The little circles show the observed percentages for the three income levels, the asterisks show the observed responses for the national sample as a whole. For no reason that we can discern the national data lie precisely on the following curve:

$$\text{Percent willing} = 98 - 14.3 \ln p.$$

This formula is shown as the middle curve in the figure and can be interpreted as a demand curve since it shows the proportion of the population that alleges that it would be willing to purchase a commodity called "clean natural environment" at various prices if it could do so as a matter of individual choice. According to this formula if the clean environment commodity were essentially free 98 percent of the population would buy it, and virtually no one would buy it at a price greater than $900 a year. The demand curves shown for the three income levels were calculated by assuming that they all had the same functional form and slope coefficient but different intercepts. The pooled sample curve lies so close to the curve for the middle income bracket that it is not feasible to show it separately. The main conclusion to be drawn is the unsurprising one that clean environment is a superior good and that at any stated price more people at higher income levels are willing to buy it than at lower incomes.

A similar question was asked in a survey two years later. Then respondents were asked how great a tax increase they would accept willingly to finance improvement in our natural surroundings. The responses are shown in Figure 4. The main impression is the same as that given by the earlier survey, but no simple family of curves could be found to represent the results. For that reason, and because the question asked in the earlier survey was more inclusive, we have used the 1969 data in our analysis.

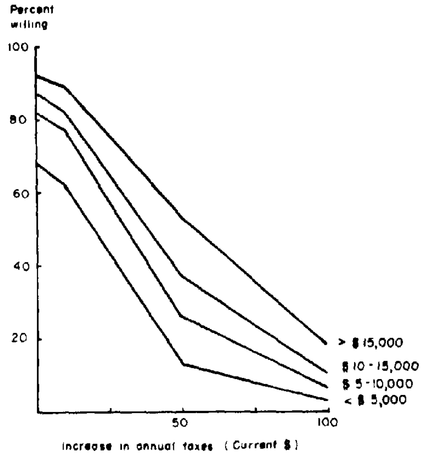


FIGURE 4. ASSERTED WILLINGNESS TO PAY TAXES FOR ENVIRONMENTAL IMPROVEMENT, 1971 SURVEY

Source: National Wildlife Federation.

Thus far we have seen estimates of the costs and the benefits of the national pollution control programs distributed by income classes. Now we wish to bring the two estimates together, but before doing so we should make it clear that we do not regard the computations to come as having any high degree of reliability. They are perhaps suggestive but certainly untrustworthy. Starting with the willingness to pay data for 1969 we can calculate the area under each of the "demand curves" and from that compute the average willingness to pay for families in the corresponding income bracket. The Public Interest Economic Center estimates, with the help of some interpolation, provide directly estimates of the average costs per family in each income bracket. These two sets of estimates are compared in Table 2 for the 1976 level of

TABLE 2—INCIDENCE OF BENEFITS AND COSTS OF POLLUTION CONTROL PROGRAM, 1976
(Current dollars)

(1) Income Class	(2) Average Willingness to Pay Per Family	(3) Average Cost Share Per Family	(4) Excess Burden Per Family
>$5,710	$ 60	$121	$ 61
$5,710 -$11,410	214	205	−9
<$11,410	608	549	−59

Sources: Col. (2) Computed from demand curves fitted to National Wildlife Federation, 1969 survey. Col. (3) interpolated from estimates of Public Interest Economics Center.

expenditure.[1] According to this table which, we repeat, should be accompanied by the Surgeon General's warning, the cost of the pollution control program to middle bracket families was just about what they would be willing to pay to obtain a clean environment. Lower bracket families, on the average, were required to pay some $60 more per family than they regarded environmental cleanup as worth, while the average burden on the upper income families was about $60 less than they said they would be willing to contribute to obtain a clean environment.

It is hard to resist believing these estimates, though we should not do so. Together with other indications they suggest very strongly that the environmental protection programs entail a redistribution of income, perhaps of substantial magnitude. But redistributing income is not a proper function for the Environmental Protection Agency. Some people object vigorously to redistributing income as an incident to government programs enacted for other purposes. We cite only Wicksell and, currently, Robert Nozick. Furthermore it is bad politics and imperils the acceptability of programs that would be desirable except for their redistributive side effects. Now, the amount and nature of the redistribution incidental to any government program are not unalterable. They can be changed by changing the techniques used to achieve the objectives and by changing the method of financing. For example, federal taxes are much more progressive than state and local taxes; to the extent that the program is implemented by federal activity or by state and local activity financed by federal subventions, the incidence of the costs will be correspondingly progressive. It appears desirable and sensible in planning programs —environmental as well as others—to give thought to ways of minimizing the amount of redistribution that they entail.

But redistributive effects are not symmetrical. There is no particular harm in charging some citizens less for programs than they would be willing to pay. The harm inheres in compelling some citizens to pay more for government programs than they would be willing to but for the coercive powers of the government. There also lies the political peril. So the thing to be minimized is the extent to which the government compels citizens to buy things that they do not want (at least at the price charged) rather than the aggregate amount of funds redistributed. We call this undesirable aspect of implicit redistribution "exaction," meaning by that term the total amount of the burden imposed on citizens in excess of the amount that they would be willing to pay for the program being financed. The amount of exaction entailed by the pollution abatement program cannot be estimated from the data in Table 2, but it must be considerable because the net excess burden on the families in the lowest income stratum is nearly $800 million a year.

A little example will help make this notion concrete and will bring out some of its implications. Suppose, then, that there are four towns who share a lake and contribute to polluting it. The State Pollution Control Board requires that the level of some pollutant in the lake be reduced by ten parts per million *(ppm)*. Any of the towns could achieve this goal individually by

[1]The rather strange income levels result from inflating incomes of $5,000 and $10,000 as of 1969 to 1976 price levels.

TABLE 3—FOUR POLLUTING TOWNS
(Hypothetical data)

Town	Cost Per Thousand Gallons Treated	Abatement Per Thousand Gallons Treated (.1 *ppm*)	Cost Per .1 *ppm* Abated	Willingness to Pay Per .1 *ppm* Abated
1	$3,000	8	$ 375	$ 80
2	4,000	10	400	40
3	3,600	6	600	120
4	4,000	4	1,000	160

treating a sufficient volume of its discharges, but the costs are widely different. The data are shown in Table 3. In Town 4, for example, the costs of treatment are $4,000 per thousand gallons and the abatement achieved is .4 *ppm* Thus a reduction of 1 *ppm* would cost Town 4 $10,000. The data for Town 1 show that they could achieve a 1 *ppm* reduction in pollution at a cost of only $3,750 The final column shows the value that each town places on a unit of pollution abatement The dice have been loaded in this example so that the towns that are situated where they can reduce pollution cheaply are also the towns that place least value on that achievement

Clearly the cheapest way to reduce pollution by 10 *ppm* is to have Town 1 do the whole job Aggregate costs will then be $37,500. But then an exaction would fall on Town 1 amounting to $29,500 ($37,500–$8,000) and they would take little comfort from the knowledge that the other three towns were receiving unpaid-for benefits of $4,000, $12,000, and $16,000 Indeed, that method of reducing pollution in the lake might seem to them, and to the courts, to be intolerably unfair.

A little arithmetic will show how very unfair that method of pollution abatement is. If Town 1 were permitted to treat five gallons less, the exaction levied on it would decline by $15, provided that some other town made up the deficiency in treatment. Town 2 could make up the deficiency by treating four gallons, at a cost of $16. This would not constitute an exaction from Town 2 but would only reduce their unpaid-for benefits. Total exactions would thereby be reduced by $15, but at the cost of increasing the aggregate costs of water treatment by $1. In short, it would cost $1 to reduce exactions by $15.

At this point an essential issue arises. Any sensible economist will ask: If exactions are to be reduced by $15, wouldn't it be better simply to have Town 2 pay Town 1 $15 (or $15.50) instead of spending $16 on its own, less efficient, treatment facilities? We are now looking at the Achille's heel of all compensation arguments. That certainly would be better if the compensatory payment could be arranged, and in this little example it is easy to think of suitable institutional arrangements. But in fact, there are often virtually insurmountable obstacles to arranging compensatory payments, particularly if the groups involved are not legal entities, as is frequently the case. If compensation cannot be arranged, then to have Town 1 sustain the entire burden is no more justifiable than any other social change in which the losers are not compensated though they could be It is Pareto-efficient for Town 1 to sustain the entire burden; it is also Pareto-efficient to have some of the other towns share. Which is desirable socially depends on the relative evaluation placed on expending resources and imposing exactions, or perhaps merely on where you happen to live.

If a $15 reduction in exactions is deemed to be worth $1 or more in resource cost, then it will be advantageous to have Town 2 join in treatment up to the point where its treatment costs are as great as the benefits that it derives from the reduction in abatement, after which Town 2 will also be subject to exactions. At this point Town 3 could join in, reducing exactions further but increasing economic resource costs.

This reasoning is traced through completely in Figure 5, which shows the entire tradeoff diagram between resources costs and exactions. At the left-end all treatment is done by Town 1 at a cost of $37,500 and an exaction of $29,500. At each corner a new town joins in, reducing exactions and increasing economic costs. At the

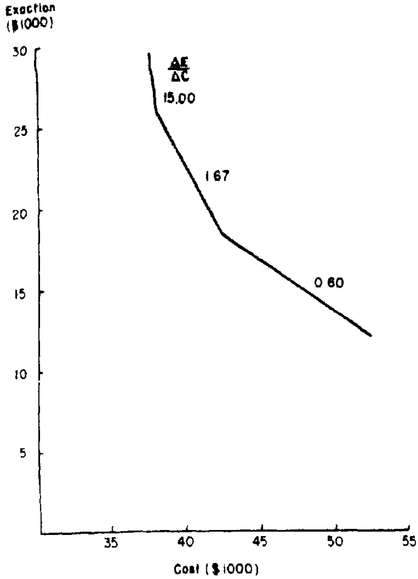


FIGURE 5 FOUR POLLUTING TOWNS, COST-EXACTION TRADEOFF

right-end all four towns treat Exactions are then as low as possible, namely $12,250, but costs have risen to $52,250 Every point on this frontier is Pareto-efficient Which should be chosen depends on the relative evaluations placed on economic efficiency and equity in this sense.

This example is so simple that the minimum possible exaction and the entire cost-exaction frontier where institutional circumstances preclude compensatory arrangements were discovered by commonsense reasoning. In general, this will not be possible. However, if the necessary data or estimates are available it is scarcely more difficult to calculate the minimum possible exaction or the cost-exaction tradeoffs than it is to make the conventional estimates of minimum cost or the benefit-cost ratio. If the situation permits the cost-minimizing plan to be found by linear programming, then the exaction-minimizing plan and the tradeoff frontier can be found by an adaptation of the linear programming algorithm. The procedure is to start with a guess as to which segments of the population will suffer exactions and which will not. Then solve the linear programming problem in which the objective is to minimize total exaction from the designated segments and the constraints are the usual ones plus the requirements that no exactions be imposed on the protected segments and that exactions from the designated segments be non-negative.

The resulting solution gives the minimum possible exaction from the designated segments, and the corresponding cost. But the initial guess might well be in error. This can be detected by inspecting the dual variables for the constraints. If, for example, the dual variable for any of the protected segments is greater than unity, that segment should be added to the exacted-from set. The reasoning is as follows. Suppose some protected segment has a dual variable of, say, 1.5 Then if an exaction of $1 be imposed on that segment, the exaction on the previously designated segments will fall by $1.50 and there will be a net decrease in exactions of $.50. Similar reasoning shows that if the dual variable for any of the segments exposed to exaction is greater than unity, that segment should be added to the protected set.[2]

If, therefore, any of the dual variables is greater than unity the initial guess should be revised by changing the designation of the segment with the largest dual variable, and the resulting linear programming problem should then be solved. It is an easy theorem that eventually a set of exacted-from segments will be found for which none of the dual variables exceeds unity, and that such a division of the population into exacted-from segments and protected segments corresponds to the minimum possible exaction. Once the minimum has been found, the tradeoff frontier can be traced by parametric programming. If the situation is

[2]An obvious generalization applies if the groups are not of equal weight or importance

such that linear programming will not suffice for the basic problem, the computations are more difficult but, in principle, the same.

Therefore, no significant conceptual or computational problems are introduced by taking exactions into account in designing and selecting government programs. This is fortunate because the extent of exactions is an important characteristic of any program, already taken into consideration by political leaders and moralists, but heretofore neglected by economists and program analysts, at least in their formal work. Program recommendations that flout this consideration face unnecessary difficulties in being adopted. Recommendations that avoid imposing unacceptable exactions are likely to have much better prognoses as well as being, indeed, more desirable socially.

As applied to the national pollution control program, concern for exactions together with the data previously introduced argue for relying as heavily as feasible on abatement measures whose ultimate incidence is progressive. In particular, this means emphasizing measures executed or financed by the federal government and imposing relatively stringent abatement requirements on polluting commodities and activities that are consumed by households in the upper income brackets. Unfortunately, there appears to be no way to avoid severe restrictions on automotive emissions. Any subsidy on automotive pollution control devices would only introduce a perverse incentive. It seems that we have to reconcile ourselves to highly regressive methods for reducing atmospheric pollution, and to redouble our efforts to compensate for them by using methods with progressive incidences for water pollution control.

Economic Growth and Climate: The Carbon Dioxide Problem

By WILLIAM D. NORDHAUS*

In contemplating the future course of economic growth in the West, scientists are divided between one group crying "wolf" and another which denies that species' existence. One persistent concern has been that man's economic activities would reach a scale where the global climate would be significantly affected. Unlike many of the wolf cries, this one, in my opinion, should be taken very seriously. The present article will first give a brief overview of the climatic implications of economic activity with special reference to carbon dioxide, and then will present possible strategies for control A more complete report with references to the literature on climatic change is contained in Nordhaus (1976).

It is thought that the economic activities which most affect climate are agriculture and energy. Of these, the latter is probably more significant, is certainly more easily analyzed, and will be discussed here. In the energy sector, emissions of carbon dioxide, particulate matter, and heat are of significance for the global climate.

I. Energy and Climate

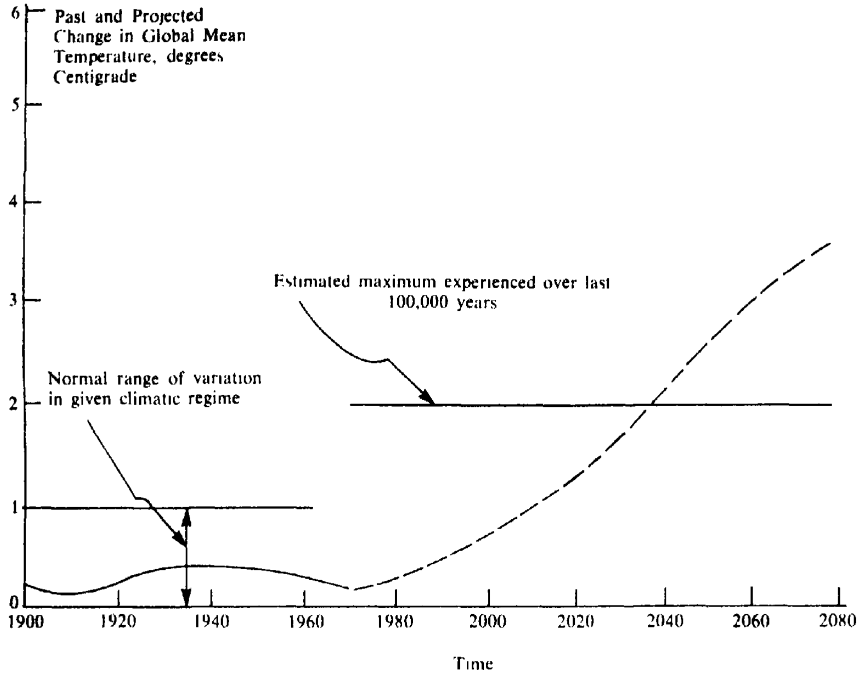
When we refer to climate, we usually are thinking of the average of characteristics of the atmosphere at different points of the earth, including the variabilities such as the diurnal and annual cycle. A more precise representation of the climate would be as a dynamic, stochastic system of equations. The probability distributions of the atmospheric characteristics is what we mean by *climate*, while a particular realization of that stochastic process is what we call the *weather*.

Recent evidence indicates that, even after several millenia, the dynamic processes which determine climate have not attained a stable equilibrium. One of the more carefully documented examples is the global mean temperature which over the last 100 years has shown a range of variation of five-year averages of about .6°C (see Figure 1).

At what point is there likely to be a significant effect of man's activities on the climate? Many climatologists feel that it is prudent to consider as significant the changes witnessed in the last century—the .6°C range. Although the estimates are uncertain, it is probable that for carbon dioxide such a change would come with an increase of approximately 20 percent in atmospheric concentrations over preindustrial levels. According to recent projections, we shall probably reach this level in the 1985–90 period (W. S. Broecker). For other sources—heat and particulates—the effects appear considerably later and are also more controversial.

A brief overview of the interaction between carbon dioxide and the climate is as follows: combustion of fossil fuels leads to emissions of carbon dioxide into the atmosphere Once in the atmosphere, the residence time appears to be very long, with approximately one-half of all industrial carbon dioxide still airborne. Because of the selective absorption of radiation, the increased atmospheric concentration is thought to lead to increased surface temperatures. The most careful study to date (S. Manabe and R. T. Wetherald) predicts that a doubling of atmospheric concentrations of carbon dioxide would eventually lead to a global mean temperature increase of 3°C. The predicted temperature increase by latitude indicates that there is considerable amplification at high latitudes. Figure 1 shows the predicted change in global mean temperature as a function of time, given the predicted emissions of carbon dioxide which we will discuss in the later part

*Cowles Foundation and Yale University



Figures up to 1970 are actual. Figures from 1970 on are projections using 1970 actual as a base and adding the estimated increase due to uncontrolled buildup of atmospheric carbon dioxide. Sources given in Nordhaus 1976).

FIGURE 1 PAST AND PROJECTED GLOBAL MEAN TEMPFRATURE, RELATIVE TO 1880-84 MEAN

of this paper. It appears that the uncontrolled path will lead to very large increases in temperature in the coming decades, taking the climate outside of any temperature pattern observed in the last 100,000 years.

II. Control Strategies

The outcome just described is the effect of an uncontrolled economy-climate system. The problem is the most extreme imaginable form of external diseconomy—one in which an individual burning a fossil fuel does not take into account the climatic consequences, and thereby affects not only the global climate, but also the climate for hundreds of years in the future.

We therefore investigate strategies for control of atmospheric carbon dioxide. A control strategy involves two aspects. On a scientific and aggregate level, the feasibility of control techniques must be explored. But there must also be a way to decentralize the controls so that nations, producers, and consumers have proper incentives to implement the control strategy on an individual level.

Figure 2 gives an overview of the model used to investigate strategies. The block labeled "energy system" can be viewed as the current system of mixed market and political mechanisms. The driving variables are energy resources, income, and population. The inter-

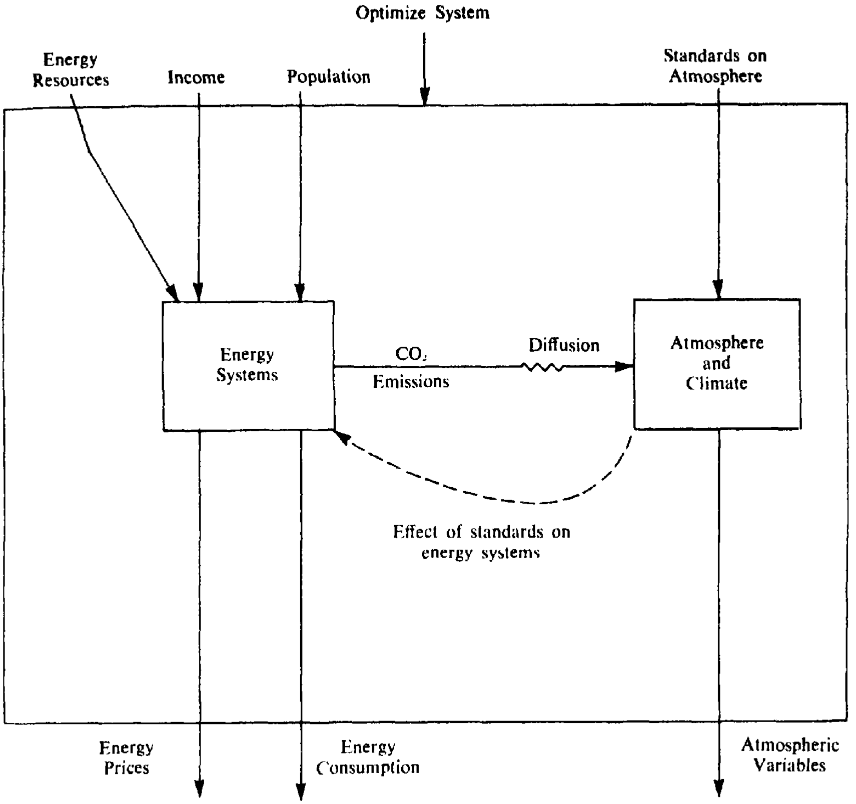


FIGURE 2 OVERVIEW OF MODEL OPTIMIZING THE ENERGY-ENVIRONMENT SYSTEM

action of supply and demand leads to a path of prices and consumption over time To account for the externalities, such as the carbon dioxide cycle, we must take into account the emissions and distribution of the effluent. This step leads us to impose standards on atmospheric concentrations, as on the righthand side of Figure 2. By imposing standards we close the loop and force the energy system to shift the composition of supply and demand. Outside the entire system there is yet another box, indicating that the entire system is being optimized.

There are two general approaches to the problem of keeping atmospheric concentrations to a reasonable level. The first strategy—the approach examined in the present paper—is to reduce emissions of carbon dioxide. This takes place basically by substituting noncarbon-based fuels for carbon-based fuels. The second strategy is to either offset the effects of emissions of carbon dioxide or use natural or industrial processes to clean out the carbon dioxide from the atmosphere *ex post*. To avoid the odor of science fiction, we have limited controls to the first strategy.

The final question in the optimization concerns the "standards" imposed in Figure 2. Unfortunately, although considerable scientific concern has been expressed about future trends in carbon dioxide concentration, there are no

attempts to suggest what might be reasonable standards. As a first approximation, however, it seems reasonable to argue that the climatic effects of carbon dioxide should be kept well within the normal range of long-term climatic variation. A doubling of the atmospheric concentration of carbon dioxide is a reasonable upper limit to impose at the present stage of knowledge. We also test the sensitivity of our results to limits of fifty percent and two hundred percent increases. We can only justify the standards set here as rough guesses and we are not certain that we have even judged the *direction* of the desired movement in carbon dioxide correctly, to say nothing of the absolute levels.

The second problem of controlling carbon dioxide is implementation on a decentralized level. Because of the externalities there are no market or political mechanisms which ensure that the appropriate level of control will be chosen. The procedure in the present paper will estimate an efficient way of allocating energy resources so as to satisfy the carbon dioxide constraint. To implement this efficient path implies that we are implicitly putting a positive price on emissions of carbon into the atmosphere, "carbon taxes," as a way of implementing the global policy on a decentralized level.

The model used to calculate the effects of imposing standards is an extension of earlier work (see Nordhaus, 1973, for a description of an early version of the energy model, and Nordhaus, 1976, for the details of the carbon model). It can be written in highly oversimplified form as follows: the energy model attempts to simulate the market allocation process. Thus, let U_{it} be the present value of the marginal utility, in terms of income, of good i in year t; c_{it} be the present value of the cost of good i in year t, both discounted at 10 percent; and let x_{it} be the level of activity. Under suitable assumptions (see Samuelson) a market allocation can be described by the mathematical programming problem:

$$(1) \qquad \underset{\{x_{it}\}}{\text{maximize}} \sum_{t=1}^{T} \sum_{i=1}^{n} [U_{it} - c_{it}] x_{it},$$

subject to constraints on the activities as well as the following resource constraints,

$$(2) \qquad \sum_{t=1}^{T} \sum_{i=1}^{n} A_{ij} x_{it} \leqq \bar{R}_j, \; j = 1, \ldots, m,$$

where A_{ij} is the content of scarce resource j per unit activity of good i, and $\bar{R}_j$ is the amount of scarce resource R_j which is available. In the actual problem examined, the goods x_{it} are composed of different energy goods (6 different fuels used in 4 different sectors), for 2 different regions of the world (United States and the rest of the world), for 10 time periods of 20 years each.

Equations (1) and (2) describe the energy system. Suppose we now wish to examine the carbon cycle as well. To do this we add a second block of equations describing the emissions and diffusion. If $\gamma(\ell\ell, i)$ is the emissions per unit activity x_{it} into stratum $\ell\ell$, emissions in a given period, $E(\ell\ell, t)$ are:

$$(3) \qquad E(\ell\ell, t) = \sum_{i=1}^{n} \gamma(\ell\ell, i) x_{it}, \qquad \ell\ell = 1, \ldots, L.$$

Next denote $M(\ell\ell, t)$ as the total mass of carbon in a given stratum at the end of period t and $D(i, j)$ as the transition probabilities of a unit mass moving from stratum i to stratum j. From the basic diffusion equations we have

$$(4) \qquad M(\ell\ell, t) = \sum_{i=1}^{L} D(i, \ell\ell) M(i, t-1) + E(\ell\ell, t), \; \ell\ell = 1, \ldots, L, \; t = 1, \ldots, T$$

Finally, we impose standards on the energy sector that the total mass in a given stratum should not exceed $St(\ell\ell)$:

$$(5) \qquad M(\ell\ell, t) \leqq St(\ell\ell), \text{ all t.}$$

To implement the controls, we add equation set (3), (4), and (5) to our original problem in (1) and (2) and solve the optimization problem.

III. Results

The first question to investigate is whether the carbon controls we have suggested are feasible. Any nonfossil fuel energy source (fission, fusion, solar, or geothermal) will be an option for meeting the carbon dioxide constraint since nonfossil fuels have no significant carbon dioxide emissions.

The second question refers to the quantities in the controlled and uncontrolled paths. Table 1 shows the calculated *U.S.* energy consumption and world carbon emissions along the uncontrolled and controlled paths. These show two surprising results. First, although the time path of emissions is severely constrained, the total energy consumption is not. In fact, in later periods (when the nonfossil fuel production becomes most significant), consumption is higher because of the lower thermal efficiency of nonfossil sources. Second, it is surprising that the effect of a carbon constraint on *current* energy consumption (and on the composition of consumption) is almost negligible; it is only in the later periods that an efficiently designed program leads to noticeable modifications of the energy system

In an optimization framework, as in an economy, constraints have their costs in terms of the objectives of the optimization, and associated with each of the carbon constraints are shadow prices on emissions. The last row of Table 1 gives the shadow prices for carbon emissions in the control program. The prices per ton start very low ($.14 per ton carbon), become significant in the third period, and rise to a very high level of around $90 a ton (1975 prices) by the end of the next century. These should be compared with the prices per ton of carbon of carbon-based fuels, which are around $25 a ton for coal, $100 a ton for petroleum, and $200 a ton for natural gas.

We can also ask what the carbon dioxide constraints are costing *in toto*. Table 2 shows the discounted cost of each of the three control programs, calculated from the attained value of the objective function in (1). Clearly the control of carbon dioxide is a very expensive operation —the middle control path 3 has discounted costs of $87 billion in 1975 prices. As a corollary, it is evident that the social return to new "carbon control technologies"—the science fiction stories referred to earlier— would be very high if carbon dioxide were to be controlled.

It should also be noted that, since at the present the only proven large-scale and low-cost alternative to carbon-based fuels is nuclear fission, the outcome of the nuclear debate will significantly affect the prospects of carbon dioxide control If nuclear fission were to be constrained along with carbon emissions, then, until major breakthroughs in alternative technologies become available, the growth rate of energy consumption would be effectively con-

TABLE 1—ENERGY CONSUMPTION, CARBON EMISSIONS, AND CARBON EMISSION TAXES

	1970 (Actual)	1980	2000	2020	2040	2100
Energy Consumption, United States, 10^{15} btu/yr						
Uncontrolled CO_2	71	76	92	155	250	395.
100 percent increase CO_2		76.	92	142.	160	405
Global Carbon Emissions, 10^9 tons/yr						
Uncontrolled CO_2	4 0	6.9	10 7	18 4	40 1	45 4
100 percent increase CO_2		6 9	10 7	16.6	16 0	4.9
Carbon Emission Tax ($/ton)						
Uncontrolled CO_2	00	.00	00	00	00	00
100 percent increase CO_2		14	1.02	8 04	67 90	87.15

Notes Carbon emissions are tons of carbon dioxide, carbon weight, while carbon taxes are calculated dual variables in the efficient program, and have the dimension of 1975 dollars per ton carbon weight of emission Source is Nordhaus (1976).

strained to zero. Preliminary estimates indicate that the cost of prohibiting nuclear power along with a limitation on carbon dioxide concentrations is around five times the most restricted case in Table 2.

In summary, an efficient program for meeting reasonable carbon dioxide standards appears feasible and, moreover, requires little change in the energy allocation for 20 to 40 years. Subject to the limitations of the techniques used here, we can be relatively optimistic about the technical feasibility of a carbon dioxide control strategy. The central question for economists, climatologists, and other scientists remains: How costly are the projected changes in (or the uncertainties about) the climate likely to be, and therefore to what level of control should we aspire? And for students of politics, the question is: How can we reasonably hope to negotiate an international control strategy among the several nations with widely divergent interests?

TABLE 2—COST OF CARBON DIOXIDE CONTROL PROGRAMS

	Path			
	1 Uncontrolled	2 200% Increase	3 100% Increase	4 50% Increase
Discounted Total Cost, Billions of 1975 Dollars	$0	$30	$87	$540
Discounted Total Cost as Percent of Discounted World *GNP*	0%	06%	12%	81%

Source. Nordhaus 1976.

REFERENCES

W. S. Broecker, "Climatic Change: Are We on the Brink of a Pronounced Global Warming?" *Science,* August 8, 1975, *189,* 460–63.

L. Machta and G. Telegados, "Climate Forecasting," in W. N. Hess, *Weather and Climate Modification,* New York 1974.

S. Manabe and R. T. Wetherald, "The Effect of Doubling CO_2 Concentration on the Climate of a General Circulation Model" *Atmospheric Sciences,* 1975, *32,* 3–15.

William D. Nordhaus, "The Allocation of Energy Resources," *Brookings Papers,* 1973, *4,* 529–70.

——, "Strategies for the Control of Carbon Dioxide," Cowles Foundation Discussion Paper, mimeo, 1976.

Paul A. Samuelson, "Market Mechanisms and Maximization," *Collected Papers,* Cambridge, Mass. 425–504

S. Schneider, *The Genesis Strategy,* New York 1976.

Externalities in a Regulated Industry: The Aircraft Noise Problem

By JEROLD B. MUSKIN AND JOHN A. SORRENTINO, JR.*

Airline noise is an externality in the traditional sense of being a byproduct of normal economic activity. It affects the population around airports in single-event bundles and cumulatively. The transient nature of noise itself is unlike most other externalities. As with the others, however, general "improvements" in technology and increases in population have made the situation more difficult to tolerate. The combination of these things has led to a critical situation for the affected population since it involves its physical and mental health.

From among the various methods of dealing with externalities, we choose the effluent charge scheme. It generally allows each firm to incorporate the environmental standard into its marginal operating choices. If each firm makes efficient decisions, then the cost to society of achieving an environmental standard will be a minimum In this paper, there are two modifications of the traditional charge scheme One is that because of the well-known difficulties in specifying and estimating social costs, we use the (estimated) direct noise abatement costs of achieving the environmental standard. The second is that airlines cannot be necessarily thought of as cost minimizers.

The charges used for the control of airline noise are generated as shadow prices in a simple linear programming model. Solution to the problem involves choosing a mix of noise-abating options that achieves proposed environmental standards at least cost. Additional bounds on the problem are limitations on service reductions and rate-of-return (*ROR*) on investment regulation. The data sources used for the program were generally fragmentary and incomplete for our purposes.

After calculating a noise charge for a hypothetical airport, we discuss the implementation and ramifications of the charge plan in the regulated airline industry. Included is a discussion of the link between noise abatement and fuel consumption

I. The Language of Aircraft Noise

Without getting into detail about how they are derived, we use two related measures of noise: 1) the *effective perceived noise level in decibels (EPNdb)* which for single noise events is a subjectively adjusted measure of noise determined by human reaction and 2) the *noise exposure forecast (NEF)* which represents the cumulative noise level during a 24 hour period. Bolt, Beranek and Newman, Inc. *(BB&N)* have shown that for aircraft class i and flight path j:

$$(1) \qquad NEF_{ij} = EPNdb_{ij} + 10\,log\,[d_{ij} + 16.67n_{ij}] - 88$$

where d_{ij}, n_{ij} are the number of daytime and nighttime flights and 88 is a normalization factor. *NEF* at a ground point is then the sum over aircraft classes and flight paths:

$$(2) \quad NEF = 10\,log \sum_i \sum_j antilog\,\frac{NEF_{ij}}{10}.$$

We estimated the affects of the adoption of noise abatement options on R, C and I via rough measures of fare, cost and investment elasticities.

Using *IBM*'s *MPS* 360 package and data

*Drexel University and Temple University, respectively

gotten from the cited references and elsewhere, we solved the *LP* problem. The package contained a sensitivity apparatus for the percent-of-goal achieved *(POG)* perturbations in the population removal goals themselves, changes in the maximum service reduction and changes in a fuel price index. The solutions were displayed for various *POG* levels (60–75 percent), fuel price indices (100–400) and the 10 percent and 20.5 percent discount rates. The solution included the optimal option bundle and total abatement costs. For the median year, 1979, we give an example *ROR* calculation using 12 percent. It was found, for example, that with a fuel index of 100, a *POG* of 65 was achieved at least cost ($822 million at *10* percent with options X_1 to X_5 and X_7 used at 100 percent and X_{10} used at 3.7 percent of full option use.

Upon perturbing the population removal constraints by 1000 persons, we obtained shadow prices to be used as noise charges. Since the amount of noise is expressed in 10 *log* equivalent terms, we expressed the charge in dollars per these units.

The disaggregation of aircraft and flight path segments by airlines allows the contribution of each airline to total noise to be determined and controlled by the charge. This extends equation (2) by summing the NEF_{ijk} over carriers k.

III. Policy Implications

Since the data was available on a national scale but charges must be imposed at each airport, we create a hypothetical three-carrier AIRPORT. We assumed that AIRPORT is represented by 5 percent of the national program. At the 10 percent discount rate and a fuel index of 100, we calculated a charge of 36¢ per 10 *log* equivalent unit. The 36¢ charge is assessed on each airline according to its monitored noise output beyond the allowed levels. The charge applied to the *NEF* 35 contour which was shown to be the "critical" contour in the *LP* program. The reception of the charge by the airlines themselves leads to a discussion of the structure of the industry.

The airline industry is a regulated oligopoly. Inefficiency is due not only to the market structure but to the solidification of this structure through regulation. The most pronounced symptoms of this inefficiency are the overcapacity problem, the excessive rate of equipment innovation and the proliferation of flights at preferred departure times. Two primary causes for these strategies are the proscription of price competition and the possibility of the *CAB* approving the introduction of competitors on "inadequately serviced" routes.

Populations exposed to equal *NEF* levels define *NEF* contours around an airport. The near-term goal of the Environmental Protection Agency *(EPA)* is to relieve all populations within *NEF* 45 to avoid long-term hearing effects. Since the *EPA* has never established any precise population removal goals, we create some for our model.

II. The LP Model

The choice variable in the linear programming model is the set of options for noise abatement. These are broken down into operational and retrofit. The operational options are: (X_1) a composite of reduced thrust takeoffs, power cutback departures, flap management approaches, higher altitude approaches, two-segment approaches and thrust reverse limitations; (X_2) preferential runways; (X_3) preferential flight paths; (X_4) night curfews; (X_5) aircraft type limitations; (X_6) aircraft weight limitations and (X_7) acquisition of land areas.

The retrofit options are: (X_8) acoustically treating the *JT3D* engine; (X_9) extending X_8 to the *JT8D* engine; (X_{10}) installing front fans and larger housings on the *JT8D* engine and (X_{11}) combination of X_8 and X_{10}.

The planning period was taken to be 1974–85, and all dollar magnitudes were discounted back to 1974 at both 10 percent and 20.5 percent discount rates. Let x_j be the percent level of operation of option X_j and c_j the discounted present value of the cost of operating x_j at the 1 percent level. The objective func-

tion is:

$$C = \sum_{j=1}^{11} c_j x_j. \quad (3)$$

Each option X_j has a specified potential reduction efficiency, a_{ij}. Let P_i be the original population and p_i the population permitted to remain in contour i. $P_i - p_i$ is the population to be removed. Once removed the population enters contour $i + 1$. The four contours i are *NEF* 45, 40, 35 and 30. The constraint may be specified as

$$\sum_{j=1}^{11} A_{i+1,j} x_j \geq \bar{p}_{i+1}, \; i = 1, 2, 3, 4 \quad (4)$$

where $A_{i+1,j} = [P_{i+1} a_{i+1,j} - P_i a_{ij}]$ is the "net migration" of contour $i + 1$, and $\bar{p}_{i+1} = [P_{i+1} - p_{i+1}]$ is the allowed residual in contour $i + 1$.

Another constraint is a maximum rate of reduction in seat-miles flown. If s_j is the percent reduction in seat-miles, $\bar{s}$ is the maximum permissible reduction, z_0 the initial number of seat-miles, then we have:

$$\sum_{j=1}^{11} s_j x_j \leqq s / z_0 \quad (5)$$

The *ROR* constraint, as defined by the Civil Aeronautics Board *(CAB)*, applies to *investment*, is a *minimum* and applies to the *industry* as a whole. It is currently set at 12 percent. If $R(x)$ is revenue, $C(x)$ operating cost and $I(x)$ investment given the choice of option bundle x, then the *ROR* constraint is:

$$\frac{R(x) - C(x)}{I(x)} \geqq V. \quad (6)$$

Airline performance can be attributed in part to three well-known behavioral theories: Oliver Williamson's expense preference effect, Harvey Leibenstein's *X-in*efficiency syndrome and the Averch-Johnson (A-J) overcapitalization effect. Williamson's thesis is that management may have goals other than cost minimization. In the airline industry, evidence of this is the proclivity for the introduction of new generation aircraft and the tendency not to diversify business investments outside of the airline industry itself. The X-inefficiency phenomenon may be characterized in the industry by imperfections in managerial performance due to lack of knowledge, insufficient incentives and/or sloth. The traditional A-J thesis must be modified with respect to the airline industry. The *ROR* is defined on investment rather than the capital stock. The *ROR* limit is a minimum rate for the industry The usual A-J behavior exhibited by individual firms is obscured If the situation exists that the industry *ROR* is near the 12 percent minimum and overinvestment causes the *ROR* to dip below it, then fare increases will generally be allowed It appears that since there is no limit on how high the *ROR* can be in this case, the industry would not be induced to hover around the minimum simply to obtain price increases. There is, however, a more subtle bound on how high *ROR* may go. When the *ROR* is high, the *CAB* is generally moved to introduce competitive carriers on the relevant routes. Hence, overinvestment to lower the apparent *ROR* may prevent carriers from facing competition. Airlines appear to regard this as a significant impetus.

Expenditures on noise abatement may also become subject to expense preference, X-inefficiency and the A-J effect. The effect of X-inefficiency would most likely be attributed to a lack of perceptiveness of firm and social needs. The other two are more difficult to separate. Airline management may have preferences for or against expenditures on environmental protection for reasons other than profits. Overinvestment in noise abatement, including overadoption of new aircraft, can either cause fare increases (at minimum *ROR*) or forestall the introduction of competitive carriers according to modified A-J behavior. Overinvestment in abatement can enhance the image of the firm which may have pecuniary (A-J) and nonpecuniary (expense preference) benefits. They still remain, however, economically inefficient.

The introduction of aircraft noise abatement

to the situation of regulated inefficiency has several dimensions. The close technological and economic regulation of the industry by the Federal Aviation Administration *(FAA)* and the *CAB* would facilitate the activities of a Noise Abatement Authority *(NAA)*. The *NAA* can use existing monitoring techniques and the clear demarcation of aircraft to attribute noise to particular airlines for charge purposes. There is also much common technological knowledge so that possible "reaction functions" to the charges might be estimated. The *NAA* must project current cost data for each year of the planning period. This will enable it to publish a tentative schedule of year-to-year charges, allowing the airlines to make long-range decisions regarding investment, scheduling, pilot training, etc.

The feedback of airline responses to the charge in terms of noise abatement affects the linear program by changing the original populations in each contour. With these new constraints, the program can be recalculated and new noise charges established. A problem with the airline responses is that we cannot generally expect cost-minimizing behavior. Hence, along with the charge proposal, we also require that the *CAB* change regulatory policies by allowing greater competition to induce airlines (at least) to minimize costs.

Society's most general goal is to maximize the overall well-being of its citizens. We do not have a precise social welfare index with all of society's variables including airline service and airline noise as elements. In lieu of placing an explicit value on the benefit or detriment society receives from these, we simply fix their levels and find the least-cost way of achieving them. One particularly useful aspect of the analytical-computational model used above is that the "meta-problem" can be dealt with through perturbations of the parameters in the "inner-problem." Although still given in terms of costs, society's broader tradeoff decisions can be explicitly seen.

The effluent charge scheme proposed in this paper can be an efficient, operational procedure if accompanied by changes in regulatory procedures by the *CAB* inducing cost minimization. The cooperation of federal and state agencies, the airline industry and local communities will increase the convergence toward a social optimum. The reflection in the exposed populations of changes in airline noise output make the *LP* problem sensitive to airline abatement activities. This allows efficient planning for capital-intensive operations. A noise abatement strategy that fails to take account of the cumulative nature of aircraft operations, the multiplicity of noise-abating options, the diversity of the social effects of noise (and its reduction) and the efficiency of the charge scheme is bound to yield a high-cost, partially effective outcome. The current approach of the *EPA* is an example.

REFERENCES

Harvey Averch and Leland L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69

Harvey Leibenstein, "Organization of Frictional Equilibria, X-Efficiency and the Rate of Innovation," *Quart. J. Econ.*, Nov. 1969, *82*, 600–23

Oliver E. Williamson, *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*, Englewood Cliffs 1964.

Bolt, Beranek and Newman, Inc., *Aircraft Noise Analysis for the Existing Air Carrier System*, Report No 2218, Project No. 118992, report to the Aviation Advisory Committee, Washington, DC, Sept. 1972.

International Business Machines, Mathematical Programming System/360, Version 2, Linear and Separable Programming —User's Manual, Program No. 360A-CO-14X, White Plains, July 1971.

U.S. Civil Aeronautics Board, Domestic Passenger Fare Investigation, Testimony of J. Maldutis, Jr., Exhibit TW-T-B, Phase 7, Docket 21866-71, 1971.

U.S. Environmental Protection Agency, "Regulation of Aircraft Noise," Draft of the Project Report, Washington, U.S. Government Printing Office, July 1974.

EXHAUSTIBLE RESOURCES

Second Best Pricing Policies for an Exhaustible Resource

By DONALD A. HANSON*

In the theory of exhaustible resources, the classical result, originally derived by Harold Hotelling, is that the scarcity rent of the resource must increase at the rate of interest. The scarcity rent is the market price of the resource less extraction costs. At the depletion time, the market price must be equal either to the zero demand price (Orris Herfindahl) or the cost of a perfect substitute (William Nordhaus), assuming no adjustment costs in switching to the substitute (Hanson). The substitute may be either a natural resource with a higher extraction cost or a backstop technology.

The Hotelling result is a price equilibrium condition in a competitive asset market (Robert Solow). It is also an efficiency condition for allocating the resource over time in a first best world.[1] However, Solow raises the possibility that constraints creating a wedge between interest rates may be important considerations in the resource allocation problem:

> Hotelling mentions, but rather pooh-poohs, the notion that market rates of interest might exceed the rate at which society would wish to discount future utilities or consumer surpluses I think a modern economist would take that possibility more seriously. It is certainly a potentially important question, because the discount rate determines the whole tilt of the equilibrium production schedule. If it is true that the market rate of interest exceeds the social rate of time preference, then scarcity rents and market prices will rise faster than they "ought to" and production will have to fall correspondingly faster along the demand curve Thus the resource will be exploited too fast and exhausted too soon. [p. 8]

In a second-best world it is not at all clear how fast the scarcity rent of the resource should increase from a social viewpoint.[2] However, for one simple case the analysis of this problem is straightforward. Suppose consumption is determined by a Keynesian consumption function with marginal propensity to consume $(1 - s)$. With consumption determined in this behavioral manner, savings may be inadequate to reduce the market interest rate to the point where it is equal to the social rate of time preference. It is argued here that for this case the scarcity rent of the resource should increase at a rate equal to a weighted combination of these two interest rates.

I. The Consumption Rate of Interest

The following discussion of the consumption rate of interest is based on the approach taken by Kenneth Arrow. Society must agree on a fair method of evaluating tradeoffs in consumption between periods. Suppose utility at time t depends only on current consumption c_t and not on consumption in other periods. Let $u(c_t)$ be a smooth, increasing, concave function. Further, suppose society wishes to discount future utilities at a constant rate ρ. These assumptions are sufficient to define the marginal rate of substitution *(MRS)* between present consumption and consumption in some future period t in the usual way holding the total discounted sum of utilities fixed. As the length between periods approaches zero, the *MRS* approaches

*NSF Energy-Related Postdoctoral Fellow, Massachusetts Institute of Technology The author wishes to thank Dagobert Brito and William Oakland for many helpful comments and Joseph Stiglitz and Martin Weitzman for discussing an earlier version of this paper. It was prepared with the support of National Science Foundation grant GK–42098.

[1]For a definition of intertemporal efficiency see Robert Dorfman, Paul Samuelson, and Solow, Ch 12

[2]This point is illustrated by the fact that the appropriate discount rate for evaluating public investment in a second best world depends on both interest rates (see Stephen Marglin).

$$(1) \qquad \frac{u'(c_t)}{u'(c_0)} e^{-\rho t}$$

Equation (1) is also called the consumption discount factor. It gives the amount of present consumption c_0 which society would be willing to sacrifice in order to increase future consumption c_t by one unit. The greater the concavity of $u(c_t)$, the more heavily the consumption discount factor depends on the relative magnitude of c_0 and c_t.

The consumption rate of interest is the proportional rate of change of (1).

$$(2) \qquad r_c = \sigma \frac{\dot{c}_t}{c_t} + \rho$$

where $\sigma = -c_t u''(c_t)/u'(c_t)$ is the elasticity of marginal utility of consumption For the class of functions:

$$u(c_t) = \begin{cases} \frac{1}{1-\beta} c_t^{1-\beta} & \beta \geq 0, \beta \neq 1 \\ \log c_t & \beta = 1 \end{cases}$$

σ is independent of c_t and is equal to β If $u(c_t)$ were more concave, σ would be greater and the growth rate of consumption would be more important in determining r_c. Note that r_c may be negative if consumption is decreasing.

II. Intertemporal Resource Allocation

Consider a natural resource which is divided into a discrete number of grades each of which is homogeneous with constant per unit extraction cost. Extraction costs differ between grades and hence the scarcity rents differ. The results derived in this section apply to the scarcity rent within a given resource grade.

If the natural resource is allocated optimally over time, in each period its value marginal product must be equal to the shadow price on the resource stock.[3] All prices are taken to be in present value terms. The present value shadow price on the resource stock of a given grade must be constant, since the stock neither grows nor depreciates over time. Let the constant be η. Suppose firms in the economy use the resource up to the point where its marginal product equals its market price p_t. Further, suppose there is a constant extraction cost "a" associated with the resource. Then the net social marginal product of the resource is $(p_t - a)$. Let V_t be the present value of an additional unit of output at t. Then the necessary condition for determining optimal resource flow can be written as

$$(3) \qquad V_t(p_t - a) = \eta$$

From (3) it is clear that the growth rate of the net price is given by

$$(4) \qquad \frac{\dot{p}_t}{p_t - a} = -\frac{\dot{V}_t}{V_t}$$

In the conventional theory V_t is also equal to the present value of an additional unit of capital at time t and hence the righthand side of (4) is the own rate of interest for capital or the rate of return on investment.

Now suppose that the marginal propensity to save is "s" and that investment equals savings. Then the value of an additional unit of output is given by

$$(5) \qquad V_t = (1 - s)u'(c_t)e^{-\rho t} + s\lambda_t$$

where the amount $(1 - s)$ is consumed and valued using the consumption discount factor (1) (with $u'(c_0)$ normalized to one) and the amount s is invested and valued at the shadow price of capital λ_t. Note that in a first-best world savings must be chosen so that

$$V_t = u'(c_t)e^{-\rho t} = \lambda_t.$$

Taking s to be constant, the growth rate $\dot{V}_t/V_t$ can be calculated based on (5). Then (4) becomes

$$(6) \qquad \frac{\dot{p}_t}{p_t - a} = (1 - s)w_t r_c + s z_t \tilde{r}_k$$

[3]Although this condition may not hold for every type of constrained optimum resource allocation, it certainly must hold for the case where the constraint is a fixed marginal propensity to save, as analyzed here

where $\bar{r}_k \equiv -\dot{\lambda}_t/\lambda_t$, the own rate of interest for capital; $w_t = u'(c_t)e^{-\rho t}/V_t$, the value of consumption c_t relative to output at t; and $z_t = \lambda_t/V_t$, the value of investment at t relative to output at t. Note that

$$(1-s)w_t + sz_t = 1$$

Hence, the growth rate of the net price of the resource is a weighted average of the consumption rate of interest and the own rate of interest for capital.

To simplify (6), suppose that capital does not depreciate.[4] Further, suppose that capital earns a rent r_k in the private market which is equal to its marginal product.[5] Then the social present value of capital is the sum over all future time τ of the marginal product of capital at τ with present value V_τ:

$$\lambda_t = \int_t^\infty V_\tau r_k \, d\tau$$

Then by differentiating λ_t it is seen that

$$\bar{r}_k \equiv -\frac{\dot{\lambda}_t}{\lambda_t} = \frac{V_t r_k}{\lambda_t} = r_k/z_t$$

and (6) becomes

$$\frac{\dot{p}_t}{p_t - a} = (1-s)w_t r_c + s r_k \tag{7}$$

where r_k is the market rate of interest. For the case where society undersaves over the planning period, which means the value of capital λ_t exceeds the value of consumption $u'(c_t)e^{-\rho t}$, then $w_t < 1$. In this case

$$\frac{\dot{p}_t}{p_t - a} < (1-s)r_c + s r_k \tag{8}$$

[4]If capital depreciates at a fixed rate, the results will still hold provided the depreciated capital is replaced using current output and s is defined as the marginal propensity to save the remaining output (net national product).

[5]A tax on income from capital which would create a wedge between the interest rate and the social rate of return on investment is not considered here.

This is the main result: if the economy is undersaving, then the optimal growth rate of the resource price should not exceed a weighted average of the consumption rate of interest and the market rate of interest according to (8). Consider a numerical example where $s = 10$ percent, $r_k = 10$ percent, $\rho = 1$ percent, $\sigma = 1$, $\dot{c}/c = 2$ percent and, hence, $r_c = 3$ percent. Then the optimal price of the resource must not grow faster than

$$(.9)\ 3\% + (.1)\ 10\% = 3.7\%$$

compared with a 10 percent growth rate in the conventional theory.

Finally, O. C. Herfindahl has shown that if the demand function for the resource is fixed, then decreasing the growth rate of the scarcity rent must increase the present price p_0, decrease present resource flow, and postpone the depletion time (see his Figure 3.2).

III. Conclusions

This paper derives the appropriate growth rate, from a social viewpoint, of the scarcity rent of a natural resource in terms of the consumption interest rate r_c and the market interest rate r_k when the marginal propensity to save s is fixed. The result does not depend explicitly on a production function relating capital and resource flow to output. However, it specifies a relationship which must hold between the marginal products of capital and resource flow where the former is equal to the market interest rate and the latter is equal to the resource price.

This relationship should be interpreted cautiously as a partial equilibrium result. For example, if savings is a social decision variable constrained only by total output and r_k is large, then the savings ratio might be chosen close to one so that capital is quickly accumulated. However, then the marginal product of capital r_k would decrease. In this case the Hotelling result holds and the growth rate of the scarcity rent would equal r_k, but r_k would be less than in the fixed savings ratio case. Also the resource price path boundary

condition at the depletion time may depend on r_k, since the cost of the substitute may depend on the availability of capital.

Other reasons for the market interest rate to exceed the social rate of time preference include taxes on income from capital, the existence of risks facing individuals or firms which are not social risks, and differing rates of discounting future utility between individuals currently living and society as a whole including future generations. Deriving the appropriate resource price path under any of these conditions is considerably more difficult than the case analyzed here.

REFERENCES

Kenneth J. Arrow, "Discounting and Public Investment Criteria," in A. V. Kneese and S. C. Smith, eds., *Water Research*, Baltimore 1966.

Robert Dorfman, Paul A. Samuelson, and Robert M. Solow, "Linear Programming and Economic Analysis," New York 1958.

Donald A. Hanson, "Competitive Price Behavior of an Exhaustible Resource Where the Rate of Substitution is Constrained," *Intern. Econ. Rev.*, Feb. 1977, forthcoming.

O. C. Herfindahl, "Depletion and Economic Theory," in M. Gaffney, ed., *Extractive Resources and Taxation*, Madison 1967.

Harold Hotelling, "The Economics of Exhaustible Resources," *J. Polit. Econ.*, April 1931, *39*, 137–75.

Stephen A. Marglin, "The Opportunity Costs of Public Investment," *Quart. J. Econ.*, May 1963, *78*, 274–89.

William D. Nordhaus, "The Allocation of Energy Resources," *Brookings Papers*, *3*, 1973, 529–76.

Robert M. Solow, "The Economics of Resources or the Resources of Economics," *Amer. Econ Rev. Proc.*, May 1974, *64*, 1–14

Public Policies Toward the Use of Scrap Materials

By Robert C. Anderson*

Numerous proposals to stimulate the flow of recycled materials have been discussed in recent sessions of Congress. The thrust of the proposals is that recycling rates are too low and that the federal government should offer incentives to aid the competitive position of the secondary materials sector. This paper examines the principal economic arguments which have been offered in support of a federal program of recycling incentives and analyzes some of the recent legislative proposals in light of available information on the structure of the secondary materials industry.

I. Background

One key argument advanced in support of recycling incentives is that tax equity should be established between recyclers and primary material producers. The depletion allowance for mineral production, expensing provisions for mineral exploration and development, and capital gains treatment of profits on standing timber all contribute to a reduced tax burden for primary material production. Some recent legislative proposals have suggested that similar subsidies be offered to recyclers (e.g., depletion deductions in *H. R.* 148).

A second argument is based upon market failure attributable to external diseconomies in primary material production (air and water pollution and the disruption of scenic natural environments). Because resource recovery would lessen these environmental damages and create few new ones of its own, one may wish to subsidize the secondary materials industry. The force of this argument has been reduced by statutes such as the National Environmental Policy Act, the Federal Water Pollution Control Act, and the Clean Air Act, all of which were designed to reduce environmental degradation.

The existing pattern of municipal subsidization of postconsumer waste disposal constitutes a deterrent to recycling. Most of the nation's households do not pay for their incremental use of waste collection and disposal services. Free disposal for individual users with disposal costs covered out of general tax revenues induces the production of greater quantities of solid residuals than if disposal fees were levied. In addition, free disposal biases the ultimate disposition of solid residuals toward disposal and against recycling. Kenneth Wertz has estimated the price elasticity of demand for municipal waste collection to be $-.15$ and calculated that imposition of disposal fees would reduce the demand for municipal waste collection by about 15 percent. The fraction of this 15 percent which would represent increased resource recovery is unknown

A final argument is that the existing structure of federal regulation favors primary production over secondary material recovery and should be balanced with incentives for recycling. Specific examples of discriminatory regulations include the General Mining Law of 1872 which grants mineral rights to those making a discovery on open federal lands, the pattern of freight rates which are alleged to favor primary over competing secondary raw materials, and labeling regulations which require source identification for many products which contain recycled materials.

Do the arguments which have been offered provide an economic rationale for the creation of

*Environment Law Institute and the University of Maryland. Fred Smith, Oscar Albrecht, Will Irwin and Roger Dower provided helpful comments. Financial assistance was provided by the U S Environmental Protection Agency

federal recycling subsidies? Here each point is examined in turn. The history of the depletion allowance, as reviewed by Talbot Page (1976a), suggests that legislative desires to promote equity among taxpayers has been the primary force behind the gradual expansion of coverage of the depletion allowance. If depletion is to be allowed for coal, iron, copper, and other minerals, as well as sand, gravel, and sea shells, why not extend it to other exhaustible sources of industrial raw materials, namely the recycling sector? Arnold Harberger has argued that efficiency in the factor market requires that income be taxed at similar rates in all activities. Thus one may argue that extension of tax subsidies to recyclers is desirable. However, this would extend the distortion between investments in material production and investments in general manufacturing. Rather than extend income tax subsidies to recyclers the existing tax subsidies for virgin material production should be eliminated if one is interested in promoting efficiency in the allocation of factors of production.

Recycling may result in less environmental disruption than does production of an equivalent output of virgin raw materials. But to subsidize recycling merely because it produces fewer harmful external effects would not promote efficiency in resource allocation. A subsidized recycling industry may compete more effectively with a polluting primary sector, but the reduction in primary material demand provides little incentive for primary producers to reduce their production of effluents.

Most consumers have free use of municipal solid waste systems, creating a divergence between private and social costs which adversely affects consumer incentives to recycle solid wastes. On the other hand, producers of industrial scrap materials must pay for solid waste disposal. A subsidy to recycling which is intended to offset the effects of free disposal should be designed so that only scrap originating within the municipal waste stream is eligible. No corrective action for industrial scrap is warranted.

Inasmuch as charging for disposal appears to be the obvious solution to this market failure, we should review the case for not pricing disposal services. One argument in favor of free disposal is that littering, another form of free disposal, will become more attractive whenever disposal fees are levied. Because littering involves large social costs, it may be desirable to subsidize disposal. The littering issue is largely unresolved, but it may constitute an important second-best argument in favor of recycling subsidies rather than the imposition of fees for use of the municipal waste system. Another argument is that the costs of administering a system of disposal fees might outweigh the benefits, although the merit of this argument may be questioned since disposal fees have an established history of use in some western cities (e.g., San Francisco).

Recapitulating the discussion to this point, it appears that offsetting free private disposal of postconsumer waste is the only valid justification for a recycling subsidy. In situations of direct discrimination against recycling (e.g., freight rates and labeling requirements), the indicated course of action would be to remove the discriminatory regulation, if not justified by the protection of other interests, rather than grant offsetting subsidies for recycling.

II. Legislative Action

Congressional interest in augmenting the flow of recycled materials has developed on many fronts. The study of freight rate discrimination was mandated by the Railroad Revitalization and Regulatory Reform Act of 1976. Loan guarantees for facilities to recover energy from solid waste were proposed in *H.R.* 1045 and tax deductions for similar facilities were proposed in *H.R.* 1046. Tax deductions or direct cash subsidies for recycling would be granted under a variety of proposals offered before Congress. In this section we focus on two such proposals.

H.R. 148 would create deductions against taxable income from recycling analogous to depletion allowances for primary raw material production. The deductions would be similar in magnitude to existing depletion deductions: 15 percent for scrap iron and steel, 15 percent for secondary copper, and 22 percent for lead.

Some recycled commodities whose virgin material counterpart does not receive a depletion deduction would be granted deductions (e.g., wastepaper at 18 percent). Like depletion deductions, the income tax deductions under *H.R.* 148 would be limited to one-half of net income.

H.R. 10612 would grant to purchasers of recyclable materials credits against income tax liabilities. The credits would apply to only that portion of materials recycled which exceeded 75 percent of a base-year figure.[1] The credit would equal 7.5 percent of price for recycled ferrous metals, 11 percent for nonferrous metals, and 10 percent for wastepaper, subject to lower and upper bounds of \$5.50 and \$8.00 for wastepaper

The remainder of this paper is devoted to the evaluation of recycling subsidies proposed in *H.R.* 148 and *H.R.* 10612. The factors which will be considered include the projected impact on the quantity of material recycled and the cost to the Treasury relative to the value of materials which are recovered. The evaluation will be done through the application of econometric models of secondary material markets which have been developed by Franklin Fisher, et al , Robert Shriner, and Robert Anderson and Richard Spiegelman It is recognized that the econometric approach is not wholly satisfactory for the analysis of permanent changes in factor costs, but as of this writing other, perhaps more satisfactory methodologies (e.g , linear programming) have yet to be applied to an entire secondary material market.

The impacts of the proposals on the quantity of scrap recycled can be derived from the elasticities reported in Table 1. For wastepaper and scrap iron and steel we assumed that demand is dependent on the supply of substitute virgin materials by incorporating the price of pig iron (in its molten state a substitute for scrap steel) and the price of pulp wood (a substitute for wastepaper) in the respective demand equations. For both copper and lead it was assumed that primary and secondary supply substitute freely in satisfying industry demand. Under these assumptions the effects of the various legislative proposals were estimated. They are summarized below in Table 2.

TABLE 1—PRICE ELASTICITIES

Material	Demand[a]	Primary Supply	Secondary Supply[b]
Iron & Steel[c]	64	—	1.1
Wastepaper	25	—	.49
Lead	21	1 0	48
Copper	86	1 67	32

[a]Demand is domestic consumption of all copper and lead, secondary iron and steel, and wastepaper

[b]Secondary supply is for industrial and postconsumer scrap steel and wastepaper, and for only the postconsumer component of secondary lead and copper flows

[c]Demand elasticity is from Anderson and Spiegelman; supply elasticities reported by Anderson and Spiegelman and Shriner essentially identical

TABLE 2—ESTIMATED INCREASES IN RECYCLING

Material	*H R* 148	*H R* 10612
Iron & steel	2 9%	3 0%
Wastepaper	1 4%	1 6%
Lead	3 6%	3 8%
Copper	2 1%	3 2%

Note These calculations are based on the assumption that profit margins are large enough to support the full deduction To the extent profit margins fall short of that which would permit the maximum deduction, the estimated impact on recycling would be reduced

The cost to the Treasury of each of the legislative proposals can be compared to the market value of the material whose recovery would be induced. Such a comparison is reported in Table 3.

How do these costs compare with the subsidy which would offset free disposal? Disposal costs currently average approximately \$26 per ton, or some 25 percent of the value of scrap iron and steel, 65 percent of the value of wastepaper, 10 percent of the value of scrap lead,

[1]It should be noted that a tax credit which applies to output in excess of some base amount will affect new entrants to the industry quite differently from existing firms in the industry. Although marginal costs for each would be the same, average costs for new entrants would be lower. This could produce incentives for firms to exit and later reenter the industry Consequently, the costs to the Treasury may be higher than the estimates presented here

TABLE 3—COST TO TREASURY AS PERCENT OF MARKET VALUES

Material	H.R. 148	H.R. 10612
Scrap iron & steel	250%	60%
Wastepaper	620%	160%
Lead	290%	70%
Copper	340%	90%

Note. Treasury costs as a percent of market value are independent of profit margins.

and two percent of the value of scrap copper. For deductions which would apply to all scrap recovered, the cost to the Treasury equals or exceeds ten times the magnitude of the uncharged disposal fee. Because most industrial scrap and much of the easily recovered post-consumer scrap would not be eligible for deductions under the proposed *H.R.* 10612, the cost to the Treasury per incremental ton recovered is less than under legislation permitting deductions for all recovered scrap. Even so, the costs of the subsidy are in excess of twice the cost of disposal for the material recovered under the subsidy.

Rather than subsidize the recovery of post-consumer waste, it has been suggested by Fred Smith, Page (1976b) and William Baumol that disposal costs could be internalized by imposing fees upon material use anywhere in the stream of materials flows. In particular, it has been suggested that these fees be collected at the manufacturing stage, primarily for ease of administration. A product charge levied on material use at the manufacturing stage would internalize the externality caused by free disposal only in the special case where consumers had no choice as to the disposition of waste material. But consumers do have a choice between municipal waste collection and collection by scrap dealers. A product charge which increases the cost of consumer goods in proportion to the costs of disposal would affect demand for consumer goods, and hence, the design practices followed by manufacturers, but it would not affect the externality which exists when municipal waste collection is a free good.

Other approaches to shaping a materials policy have also received attention recently, particularly loan guarantees for recycling facilities, governmental stockpiling to stabilize supply and demand for secondary materials, and the creation of futures markets for secondary materials to reduce price uncertainty. Although these latter policies would not operate to correct the source of market failure, they, as well as the other policies which have been reviewed, may be deemed socially desirable in the broader context of a national materials policy—a policy which must consider explicitly such problems as the balance of payments, national security, and the resource needs of future generations.

REFERENCES

Robert Anderson, and Richard Spiegelman, "Tax Policy and Secondary Material Use," *J. Env. Econ. Mgt.*, forthcoming 1977.

William Baumol, "Statement before the Panel on Materials Policy," Senate Public Works Committee, May 1976.

Franklin Fisher, Paul Cootner and Martin Baily, "An Econometric Model of the World Copper Industry," *Bell J. Econ. Mgt. Science*, Autumn 1972, *3*, 568–609.

Arnold Harberger, "Efficiency Effects of Taxes on Income from Capital," in *Effects of the Corporate Income Tax* (Kryzaniak, ed.), Detroit 1966.

Talbot Page, *An Economic Basis for Materials Policy*, Baltimore 1976a.

———, "Statement before the Panel on Materials Policy," Senate Public Works Committee, May 1976b.

Robert Shriner, "An Econometric Analysis of Supply and Demand for Scrap Steel," presented at the meetings of Regional Science Association, Chicago, Nov. 1974.

Fred Smith, "The Disposal Charge Concept," mimeo, Sept. 23, 1974.

Kenneth Wertz, "Economic Factors Influencing Households' Production of Refuse," *J. Env. Econ. Mgt.*, Apr. 1976, *2*, 263–72.

INNOVATION AND INVENTION

Consumer Protection Regulation in Ethical Drugs

By HENRY G. GRABOWSKI AND JOHN M. VERNON*

A number of studies by economists have emphasized that government regulation often produces undesirable or unintended side effects. In this paper, we examine some effects of this nature on the structure of innovation in the pharmaceutical industry.

In the first section of the paper, we review recent changes in the regulatory environment in ethical drugs and show that they have been a major factor leading to higher costs and risks in pharmaceutical innovation. In the second section, we show that significant shifts have also occurred in the structure of innovation in this industry Namely, innovation has become more concentrated in large multinational drug firms. These firms are apparently in a better financial position to deal with the higher costs and risks of innovation and also can shift resources on a worldwide basis to offset some of the adverse impacts of regulations in this country. Some evidence concerning these international transfers is presented in last part of the paper

I. The Effects of Regulation in Ethical Drugs on the Costs and Risks of Innovation

In 1938, with the passage of the Food, Drug and Cosmetic Act, Congress authorized the Food and Drug Administration *(FDA)* to perform a premarket safety review of all new drug compounds. Despite these new regulatory controls, innovation in ethical drugs flourished over the next two decades. Several notable therapeutic advances were achieved in antibiotics, psychotropic medicines and other fields. Furthermore, drug industry *R & D* expenditures increased dramatically along with the annual volume of new chemical entities (*NCEs*) introduced commercially. While the premarket safety reviews of the *FDA* obviously resulted in time lags for all drugs and deterred some new drugs from the marketplace, regulatory review times were still quite short (7 months on average) and the annual volume of *NCE* introductions was at record levels (over 50 per year) at the end of the decade of the 1950's. (See Grabowski, 1976, Ch. II.)

In the early 1960's, following the thalidomide tragedy, *FDA* regulation of ethical drugs became much more stringent in character. A major factor in this regard was the passage by Congress in 1962 of the Kefauver-Harris Amendments to the Food, Drug and Cosmetic Act. This new law required firms to demonstrate the efficacy as well as safety of all new drugs to the *FDA* and also imposed regulatory controls on the clinical research process and on drug advertising and labeling.

One would expect the more stringent regulatory environment that evolved after 1962 to have some adverse effects on costs, risks and development times of new drug innovation. In fact, a number of studies have indicated that significant shifts took place in the economics of new product innovation in ethical drugs in the post-amendment period. In particular, studies by V. A. Mund, L. H. Sarett and others indicate that development costs and times increased severalfold after 1962. By the early 1970's, Sarett estimated that the introduction of an *NCE* required more than ten million dollars in development costs and a gestation period of 8 to 10 years in length. In addition, data developed by

*Professors of Economics, Duke University. This research was supported by a grant from the National Science Foundation, Division of Policy Research and Analysis.

W. Wardell and L. Lasagna indicate a high attrition rate on new drug candidates in the postamendment period. This is reflected in the fact that less than ten percent of the drugs entering clinical testing on humans after 1962 have become commercially available drugs. These adverse developments on the input side have been accompanied by a sizeable decline in the annual rate of *NCE* introductions in the postamendment period. (See Table 1.)

While there is little argument that innovational activity in ethical drugs has been characterized by significant adverse structural trends, there has been considerable debate about the role of regulation in explaining this situation Previous studies by Martin Baily and Sam Peltzman indicate that the 1962 Amendments had a strong negative effect on the rate of drug innovation. However, their analyses have been criticized by the *FDA* and others for not adequately discriminating between the impacts of regulation and other factors (see the discussion in Grabowski, 1976). An alternative hypothesis advanced in the literature is that a "depletion of research opportunities" has occurred in ethical drugs as a result of the rapid rate of innovation in the earlier postwar period; and that this has produced the adverse trends attributed to regulation

In a recently completed study, we have attempted to disentangle the effects of regulation from nonregulatory factors like research depletion, through a comparative international study of the United States and the United Kingdom (Grabowski, Vernon and L Thomas, 1976). International comparative analyses would seem to offer one of the most promising methodological approaches for analyzing this question. This is because a depletion in basic research opportunities influences innovational activities in all countries in a common way, whereas regulatory procedures have differed considerably across countries. This type of analysis therefore offers one of the closest things available to a natural experiment for distinguishing between these two hypotheses. Of course one must also recognize the multinational character of the firms in this industry in structuring this type of comparative international analysis.

Our comparison of the United States and *U.K.* focuses on the number of *NCE*s discovered and developed in each country, per dollar of *R & D* investment, in both the pre- and postamendment period. We found both countries experienced significant increases in the total *R & D* investment expenditures necessary to produce an *NCE* in the postamendment period. However the increase was relatively greater in the United States, where regulatory controls were much more extensive. On the basis of a production function analysis using these data, we estimated that increased regulation, by itself, roughly doubled the cost of producing and introducing an *NCE* in the United States in the postamendment period.

In summary, our analysis (along with several other studies) points to increased regulation as an important factor underlying the higher costs and risks of drug innovation in the United States.

II. Structural Changes in Drug Innovation

In this section we examine various supply side shifts and structural changes that have occurred as an apparent consequence of the much higher costs and risks of drug innovation in the United States (see also Grabowski and Vernon).

A. Innovation and Firm Size

The first issue we consider is whether innovation has become more concentrated in fewer and larger firms. Some data on this question are presented in Table 1 The first two rows show the total number of *NCE*s and the number of firms having at least one *NCE* over three successive five-year periods, 1957 to 1961, 1962 to 1966, and 1967 to 1971. These data clearly show that the number of independent sources of new drug introduction has declined significantly over time, along with the rate of total introductions.

The third row of Table 1 gives the dollar value of "innovational output" in each period. This is the total number of *NCE*s introduced in each period, weighted by their sales during the first three years after introduction. This measure of innovation, like the simple count of *NCE*s, also shows a significant downward movement over time. Table 1 next presents 4-firm and 8-firm concentration ratios of innovational output. These data indicate that the leading

innovative firms have been accounting for increasing percentages of total innovation in successive periods, and reinforce the point that the number of independent sources of innovation is declining.

The final question considered in Table 1 is whether innovation has become more concentrated in the largest drug firms. The last two rows show the share of innovational output and the share of total drug sales accounted for by the four largest drug firms (ranked by ethical drug sales) for each of these 5-year periods. Thus, in the preamendment period, 1957–61, and in the first postamendment period, 1965–66, the largest four firms accounted for a roughly equal amount of innovational output and sales. In the final period, however, the four largest firms accounted for 48.7 percent of innovational output, which was much greater than their share of sales (26.1 percent).

TABLE 1—CONCENTRATION OF INNOVATIONAL OUTPUT IN THE U.S. ETHICAL DRUG INDUSTRY

	Periods		
	1957–61	1962–66	1967–71
(1) Total Number of New Chemical Entities (*NCE*'s)	233	93	76
(2) Number of Firms Having an *NCE*	51	34	23
(3) Total Innovational Output[a] (millions $)	$1,220.3	$738.6	$726.8
(4) Concentration Ratios of Innovational Output			
4-firm	46.2	54.6	61.0
8-firm	71.2	78.9	81.5
(5) Four Largest Firms' Share of Innovational Output	24.0	25.0	48.7
(6) Four Largest Firms' Share of Total Sales	26.5	24.0	26.1

Sources: List of new chemical entities obtained from Paul de Haen *Annual New Product Parade*, various issues, all data on ethical drug sales from Intercontinental Medical Statistics.

[a]Innovational output is measured as new chemical entity sales during the first three full years after product introduction

These findings were also consistent with a polynomial regression analysis of innovational output on sales for 51 drug firms for the three periods. In the first two periods, a linear relationship between innovational output and sales offered the best statistical fit whereas in the third period a cubic relation offered the best fit, with innovational output increasing at an increasing rate over the upper range of size. Two regressions from our analysis are given below. Equation (1) is the linear regression for the preamendment period and equation (2) is the cubic regression for the most recent period.

1957–61:

$$(1)\quad \underset{(.07)}{Y = 359.35} + \underset{(7.11)}{.74\,S}, \quad R^2/F = .51/50.52$$

1967–71:

$$(2)\quad Y = \underset{(1.67)}{-11467} + \underset{(3.17)}{.94\,S} - \underset{(3.19)}{.88 \times 10^{-5}S^2} + \underset{(3.81)}{.25 \times 10^{-10}S^3}, \quad R^2/F = .64/20.7$$

where: Y = innovational output (\$000); S = total ethical drug sales in middle year of period (\$000); and t-statistics are in parentheses

It is interesting to note that the cubic regression equation in the 1967–71 period contributed .19 incrementally to R^2 compared with a linear regression and .11 compared with a quadratic regression.

The hypothesis that the largest firms in an industry will be the dominant sources of innovation dates back to Joseph Schumpeter's pioneering analysis. However, most empirical studies (including those for the drug industry) have not provided much support for the Schumpeterian hypothesis. Nevertheless, the results reported here are quite consistent with the trends in pharmaceutical innovation discussed in the first section. Given the much higher costs and risks of drug innovation in the postamendment period, it is plausible that the structure of innovation would shift in the direction of the

Schumpeterian hypothesis.

B. Innovation and the Multinational Activities of Pharmaceutical Firms

The most innovative firms in the ethical drug industry are not only relatively large in terms of domestic sales, but also tend to have a strong multinational character For example, the eight leading innovative firms in the 1967–71 subperiod (which accounted for over 80 percent of innovative output in that period) have a strong multinational orientation. Each of these firms had manufacturing plants in at least eight foreign countries, and seven of them has foreign sales in excess of 100 million dollars in 1970. While past studies of the Schumpeterian hypothesis have not considered this aspect of firm structure, it would appear to be highly relevant in the current context.

Multinational firms have some significant advantages in their ability to respond to the more stringent regulatory conditions that have evolved in this country. First, they can introduce new drug products into foreign markets (where regulatory conditions are less stringent) prior to (or in lieu of) introduction in the United States. This allows them to gain knowledge and realize sales revenues while a new drug compound remains under regulatory review and development in this country. While a firm with no foreign operations could in principle do the same thing through licensing, significant information and transaction costs exist in this situation to reduce the gains from a licensing arrangement.

In addition, multinational firms also can perform *R & D* activities in foreign countries in order to reduce time delays and the overall costs of developing new products. Some important institutional barriers do exist to this strategy however. Historically, the *FDA* has been unwilling to accept data from foreign clinical trials or patient experiences. Because of this, *U.S.* firms have incentives to perform their *R & D* in this country, even if they choose to introduce their new drugs first and in greater numbers abroad. Nevertheless, it should be borne in mind that only a small fraction of compounds entering clinical testing in the United States ever become commercial products (as noted above, Wardell and Lasagna indicate this fraction is now less than 10 percent). Multinational firms therefore have the option of screening new drugs abroad and performing duplicate *U.S.* trials on the relatively small fraction of drugs for which New Drug Applications are submitted to the *FDA*. They also can perform different phases of development alternatively here and abroad in order to reduce regulatory lags and bottlenecks.

Some descriptive statistics serve to illustrate the extensive shifts that have occurred in the behavior of multinational firms with respect to foreign introductions and clinical testing over the postamendment period. In Table 2, data on all *U.S.* discovered drugs introduced in the United Kingdom over the period 1960–1974 have been assembled in order to consider whether *U.S.* discoveries are now being introduced there before here. A *U.S.* discovered drug is defined as one originating in a *U.S.* laboratory.

Table 2 shows that in the early 1960's, the vast majority of *U.S.* discovered *NCE*s introductions in the *U.K.* become available there only after here. However, a rather dramatic shift in this situation has occurred over time. By the final subperiod, 1972–74, approximately two-thirds of the United States discovered *NCE* introductions in the *U.K.* were either introduced later, or have yet to become available, in the United States. Preliminary analysis of data on France and Germany suggest similar patterns.

The shift in firm behavior depicted in Table 2 would seem to be strongly tied to regulatory differences in these countries. We might also point out that the *U.S.* firms share of *U.K.* total ethical drug and new product sales declined in the post-1962 period (Grabowski and Vernon), thus amplifying the incentives operating on firms to modify their traditional practices of introducing new products abroad only after *U.S.* introduction.

It would seem important to note that the behavior of pharmaceutical firms in recent years represents a significant departure from the pre-

TABLE 2—INTRODUCTION OF U.S. DISCOVERED DRUGS IN THE UNITED KINGDOM, 1960–74

Period	Number of *NCE* Introductions in *U.K.* of *U.S.* Origin[a]	Number (Percent) of these *U.S.* Discovered *NCEs*:			
		In *U.S* Before	In *U.S.* Same Year	In *U.S.* Later	Not In *U.S.*
1960–62	57	38 (66 6)	13 (22.8)	5 (8.7)	1 (1.8)
1963–65	33	16 (48 4)	5 (15 1)	10 (30.3)	2 (6 1)
1966–68	24	10 (41.6)	4 (16.7)	8 (33.3)	2 (8.3)
1969–71	21	9 (42 8)	4 (19.0)	3 (14.2)	5 (23.8)
1972–74	28	8 (28.5)	2 (7 2)	6 (21.4)	12 (42 9)

Sources: Information on *NCE* introductions in the United Kingdom and the origin of each *NCE* introduction were obtained from data compiled by Paul de Haen, Inc., and the National Economic Development Office of Great Britain In cases of conflict between these two sources on the country of origin, the drug was not included in the above sample of U.S discovered introductions

[a]Drugs of *U S* origin defined as an *NCE* discovered in *U S* research laboratory

dictions of the product life cycle trade theory proposed by Raymond Vernon and others. Not only are these new drug innovations being introduced first in foreign countries with much smaller markets than the United States, but they must also be produced in their initial stages of product life in foreign plants as well. This is because *U.S.* regulatory law prohibits drugs not yet cleared by *U.S.* authorities from being exported to foreign countries. Indeed, this provision of the law would appear to provide substantial incentives for direct foreign investment by *U.S.* firms.

Data recently developed by Lasagna and Wardell also suggest some significant shifts have taken place in the location of clinical testing by *U.S.* firms. They have recently completed a study of the new drug compounds clinically tested by 15 large *U.S.* ethical drug firms over the period 1960 to 1974. (These firms accounted for 80 percent of *R & D* expenditures in the United States.) Their results suggest an increasing tendency for *U.S.* firms to perform clinical testing of new drug compounds first in foreign locations. Specifically, they found that in 1974 these firms clinically tested approximately one-half of all their new drug compounds first abroad, whereas before 1966, they performed virtually all their clinical testing first in the United States. Although industry *R & D* expenditure data indicate that the percentage of total *R & D* outlays expended in foreign countries by *U.S.* firms is still small (15.4 percent in 1974), foreign outlays are growing much more rapidly than domestic expenditures and this percentage has doubled in the space of a few years (Grabowski, Ch. III).

In summary, the data analyzed in this section indicate that *U.S.* based multinational firms are increasingly testing and marketing new chemical entities abroad before the United States. As discussed above, the option to engage in such foreign activities offers multinational firms significant advantages in dealing with the more stringent regulatory situation that has evolved in this country. It is therefore perhaps not surprising that large multinational firms now account for such a dominant share of innovation in the *U.S.* ethical drug industry.

III. Summary and Conclusions

Our results indicate that *FDA* regulation of ethical drugs has had some significant adverse effects on the structure of pharmaceutical inno-

vation. In effect, the higher costs and risks of drug innovation in the more stringent post-1962 regulatory environment have operated as a barrier to competition through new product introduction Consequently, the supply of new drugs has not only declined, but it has also become more concentrated over time in the larger multinational firms better able to deal with this more stringent environment Given the rapid spread of health and safety regulation controls throughout all sectors of the economy, further attention to the adverse effects of regulation on industry competitive structure would seem highly desirable. They constitute a potentially important source of long-run indirect costs to society that must be weighed against the benefits of these new regulatory controls.

REFERENCES

Martin N. Baily, "Research and Development Costs and Returns: The *U S* Pharmaceutical Industry," *J. Polit. Econ.*, Jan. 1972, *80*, 70–85.

Henry Grabowski, *Drug Regulation and Innovation: Empirical Evidence and Policy Options*, American Enterprise Institute, Washington 1976.

______ **and John Vernon,** "Structural Effects of Regulation on Innovation in the Ethical Drug Industry," in R. T. Masson and P. Qualls, eds., *Essays On Industrial Organization in Honor of Joe S. Bain*, Cambridge 1976, 181–206.

______, ______ **and L. Thomas,** "Estimating the Effects of Regulation on Innovation: An International Comparative Analysis of the Pharmaceutical Industry," Duke University Department of Economics Discussion Paper, Sept. 1976.

L. Lasagna and W. Wardell, "The Rate of New Drug Discovery," in Robert B. Helms, ed., *Drug Development and Marketing*, Washington 1975, 155–64.

V. A. Mund, "The Return on Investment of the Innovative Pharmaceutical Firm," in J. A. Cooper, ed., *The Economics of Drug Innovation*, Washington 1970.

Sam Peltzman, "An Evaluation of Consumer Protection Legislation. The 1962 Drug Amendments," *J Polit. Econ*, Sept. 1973, 81, 1049–91.

L. H. Sarett, "FDA Regulations and their Influence on Future R and D," *Research Management*, Mar. 1974, *27*, 18–20

Raymond Vernon, *Sovereignty at Bay*, New York 1971

The Characteristics of Optimum Inventions: An Isotech Approach

By ROGER A. MCCAIN*

Technical "progress" has become a controversial social issue, but the nature of technical progress is badly understood. Economists may best contribute to the discussion by analyzing technical change as an instance of choice subject to a constraint of limited technological opportunity.

The innovation possibility frontier (Charles Kennedy) is one hypothetical constraint on technological opportunity. Models based on the innovation possibility frontier customarily assume steady growth (E M. Drandrakis and Edmund S Phelps, William Fellner and McCain) and are rather well understood. They have been incisively criticized by Nordhaus, who proposed, as a general alternative, the hypothesis of an isotech map. An exploration of the characteristics of optimum inventions, in terms of the isotech hypothesis seems of some interest.

A single isotech, the C-isotech, is the set of all techniques attainable at a given cost, C Thus in the standard neoclassical model, the production function is the zero-isotech The isotech map will depend on the history of technical development as a whole and so cannot be stable over time.[1] The generality of the isotech hypothesis makes it possible to raise some questions of considerable interest, which are beyond the range of the innovation possibility frontier hypothesis. Because scale is a major determinant of the social impact of technology, (E. Schumaker) we shall as an example explore John K. Galbraith's "imperatives" of large scale.[2]

We first explore some characteristics of optimum inventions. We suppose that an invention is characterized by capital intensity, k, and labor intensity, n, as usual; and also by the minimum capital scale ℓ, and the durability of the capital good, m. The capital good is supposed to be a one-hoss shay. The isotech map is represented by a cost function

$$(1) \qquad \phi = \phi(k, \ell, m, n);$$

The output scale of an individual plant is ℓ/k; the employment scale is $\ell(n/k)$. We assume that a design is to be prepared in order to produce a predetermined flow of output, q, per year. The net value of the design is then

$$(2) \qquad V = \int_0^m e^{-rt}(p(t)q - w(t)nq)dt - kq - \phi$$

where $p(t)$ is the price of the output at time t, $w(t)$ the wage at time t, and r the rate of discount Writing

$$(3) \qquad v = \int_0^m e^{-rt}(p(t) - w(t)n)dt - k$$

equation (2) is

$$(4) \qquad V = qv - \phi.$$

For diagnostic purposes we maximize qv subject to a developmental expenditure constraint

$$(5) \qquad \phi \leq \bar{\phi}$$

*Associate Professor, Temple University I am indebted to participants in the *RES* Specialist Conference on Government and Innovation, to Richard Horne of the City College, City University of New York, and to Howard Teasley of the Oregon State Department of Highways, for useful discussions of some of the topics explored in this paper.

[1]This follows Nordhaus

[2]Mathematical details are available from the author

and the necessary conditions for a maximum are[2]

$$\text{(6a)} \qquad \frac{\dfrac{\partial \phi}{\partial n}}{\dfrac{\partial \phi}{\partial k}} = -W = -\int_{o}^{m} e^{-rt} w(t)\,dt$$

$$\text{(6b)} \qquad \frac{\dfrac{\partial \phi}{\partial m}}{\dfrac{\partial \phi}{\partial k}} = e^{-rm}[p(m) - w(m)n]$$

$$\text{(6c)} \qquad \frac{\dfrac{\partial \phi}{\partial \ell}}{\dfrac{\partial \phi}{\partial k}} = 0$$

These conditions are displayed in cross section in Figures 1, 2, and 3. In Figure 1, lines xx', yy', zz', have the slope equal to $-W$, where W is the discounted present value of expected wage disbursements over the life of the plant. Notice that a rise in $w(t)$ for some periods, with $w(t)$ unchanged in other periods, will raise W and so lead to a substitution of capital for labor along the constraint isotech, that is, the $\bar{\phi}$-isotech. Notice too that a smaller r would imply a larger W *ceteris paribus* and would also induce substitution of capital for labor, as we would expect. As the diagram is drawn, however, we see a larger research effort (larger $\bar{\phi}$) associated with larger savings of both capital and labor. This in part reflects the partial-analytic approach in use here; that is, it reflects the constancy of W regardless of the magnitude of $\bar{\phi}$. If we consider the aggregate economy as a whole, we might expect a larger development enterprise, $\bar{\phi}$, to be correlated with a larger W reflecting a more rapid rate of increase of $w(t)$. Figure 4 is drawn to illustrate this possibility. In Figure 4, W rises just rapidly enough with $\bar{\phi}$ so that technical progress is Harrod-neutral (k is constant provided r is), a familiar condition of steady growth.

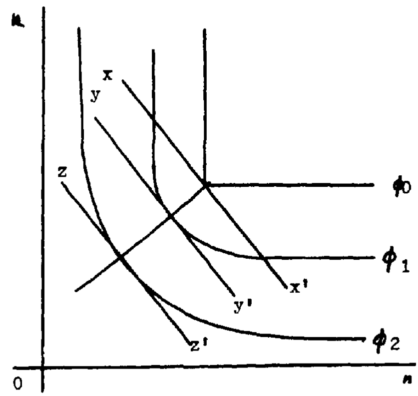


FIGURE 1

In Figure 2, curves aa', bb' and cc' have a slope of $e^{-rm}[p(m) - w(m)n]$. The downward convexity of the curves aa', bb' and cc' is most plausible, since a larger m will mean not only a smaller e^{-rm} but also (with rising $w(t)$) a smaller last period mark-up $p(m) - w(m)n$. Notice that nevertheless a smaller r will, *ceteris paribus*, entail a larger slope $e^{-rm}(p(m) - w(m)n)$ and so a substitution of durability, m, for initial capital outlay, k, along the constraint isotech $\bar{\phi}$. Notice that "technological obsolescence" would require that $p(m) - w(m)n = 0$, but this will not ever be so if $\dfrac{\partial \phi}{\partial m} \neq 0$.

Condition (6c) indicates that no value is placed on scale per se, within the relevant range (i.e., assuming ℓ is much smaller than q). Thus the slope of jj', hh', and gg' is zero, as indicated by condition (6c). If the corresponding optimum ℓ/k should be greater than q, this result would of course be invalid, and a more complex result must be invoked.

The determinants of the scale of a representative plant deserve more careful consideration. The scale of the individual plant will be an important determinant of the social and environmental impacts of a technology. However, the determinants of the scale of an individual plant have been little explored in

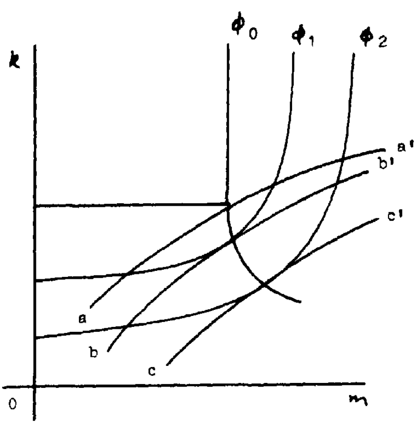


FIGURE 2

economics. Galbraith, an exception to the general neglect, assures us that the "imperatives" of technology require large scale.

What are "imperatives of technology?" If we regard technology not as an object of choice but as a given, as Galbraith often seems to do, then the "imperatives of technology" are straightforward enough: every innovation has its time, its place, and its scale, and only one scale is possible. If, however, we regard technology as an object of human choice, this interpretation becomes meaningless noise. Still, we can interpret Galbraith's hypothesis as meaning that technological opportunities are not neutral with respect to change in scale, as shown in Figure 3. Instead, we may suppose that technological opportunities are strongly biased toward large scale, as shown in Figure 5. The term "imperatives" of technology seems misplaced here, though; small scale is attainable even if at some cost.

In the Galbraithian system, the substitution of large-scale, high-productivity technology for smaller-scale technology enforces oligopoly or monopoly. (Imperfect competition then allows gains from technical development to be more nearly captured by the firm, and so leads to more research, which compounds the tendency to large scale.[3]) As an alternative to

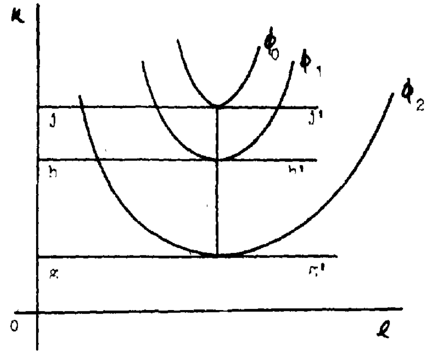


FIGURE 3

the Galbraithian scenario, we shall observe that the cause-and-effect relationship may run in the other direction. Industrial concentration may lead toward large-scale technology even at a sacrifice in terms of labor productivity, even when technological opportunities are not biased toward increasing scale, i.e., when they are as in Figure 3.

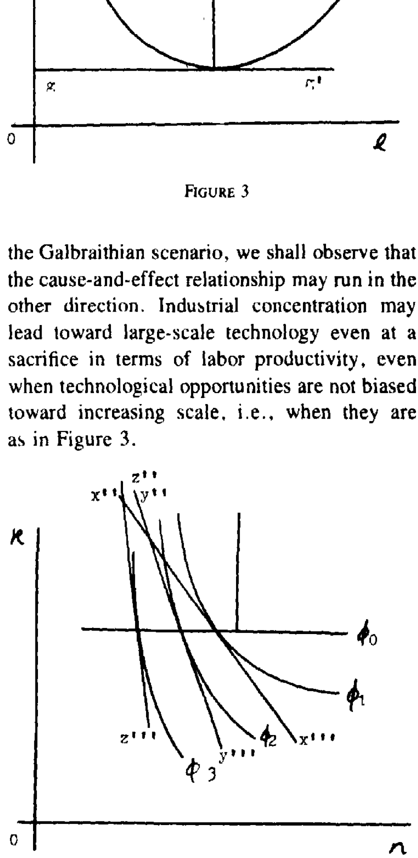


FIGURE 4

We suppose that a large firm, with smaller competitors or potential competitors, is develop-

[3]I am indebted to David Martin of the State University of New York-Geneseo, for clarifying this scenario to me.

ing new technology for its own use. For one reason or another, patents are not effective in preventing imitation of the new design by the competitors or potential competitors. Thus, the demand curve of the technology-developing firm will be

$$(8) \qquad q_f = f'(p, t); \ \frac{\partial q_f}{\partial t} < o$$

(ceteris paribus)[4] since the new techniques will be used by imitators. The imitators will attain lower cost levels than the competition now does and will presumably respond in part by expanding their output. Moreover the number of imitators will increase with time However, the vigor of imitative competition will vary with the parameters of the new design, so that more exactly

$$(9) \qquad q_f = f''\ (p, t, k, \partial, m, n).$$

In particular, we may suppose that a larger minimum capital scale, ℓ, would imply less imitative competition (William J Baumol) and so

$$(10) \qquad q_f = f'''(p, t, \ell); \ \frac{\partial f'''}{\partial \ell} < o.$$

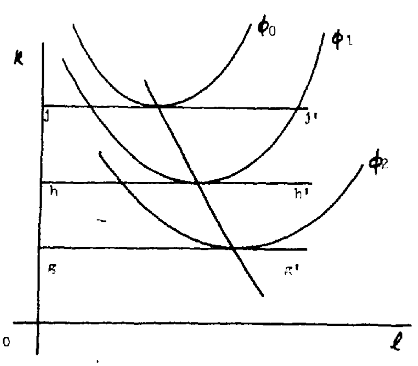


FIGURE 5

It will be more convenient to treat q_f as an independent variable, so in place of (10) we substitute[5]

$$(11) \qquad p = f(q_f, t, \ell); \ \frac{\partial f}{\partial \ell} < o.$$

The necessary condition for a maximum which corresponds to 6c is now

$$(12) \qquad \frac{\dfrac{\partial \phi}{\partial \ell}}{\dfrac{\partial \phi}{\partial k}} = -\int_o^m e^{-rt} \frac{\partial f}{\partial e} dt > o.$$

This is shown in Figure 6, where the slopes of curves jj', hh', and gg' represent the right-hand side of (12). These curves are drawn convex downward on the plausible assumption that the marginal impact of scale on price diminishes with increasing scale.

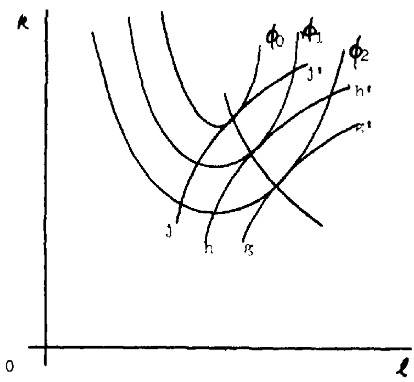


FIGURE 6

The situation would be very different if the firm were designing new equipment to sell to other firms, rather than for its own use, as, for example, agricultural implement companies

[4]This may be offset by secularly increasing demand, which is neglected here

[5]The assumption here is that the plant, once built, will be operated at full capacity This assumption allows us to avoid a problem in the calculus of variations, and can be relaxed without modifying the results Details are available from the author on request.

would design new equipment for use by farmers. Still assuming that large scale is a barrier to rapid diffusion of the new technology, the designer would then choose a smaller scale than that which would maximize productivity, given development expenditure, rather than greater. This suggests that the characteristics of optimum inventions will depend in complex ways on the characteristics of particular markets, and no doubt this is the case.

This distinction is of some importance, for it suggests that the dispute between proponents of the Galbraithian hypothesis of "technological imperatives" and of the opposite hypothesis may be resolved empirically. On the alternative hypothesis, the tendency of scale of new machinery and its correlation with the magnitude of the developmental effort will depend on whether the machinery is produced for the firm's own use or for sale in a broad market. The Galbraithian hypothesis entails no such dependency.

Technical development is neither a juggernaut nor an arena of unrestricted choice. Rather it is an arena of constrained choice. Constrained choice is a very familiar concept in economics, but the difficulty in treating technical development as an area of constrained choice has lain in the appropriate conception of technological opportunities. This paper has explored the characteristics of optimum inventions in the light of Nordhaus' conception of an "isotech map," considered as the constraint of technological opportunity. Shortcomings of this approach are its partial-analytic character and its character as a "snapshot" of technological opportunities at an instant of time. This means both that the isotech map is not stable over time and that it depends on the historical time path of past technical progress, which means in turn that the isotechs may not have the convexity which would assure uniqueness of the local optimum. Nonetheless, some insights have been gained on the economics of technology, including both extensions of familiar "neoclassical" propositions and systematization of some Galbraithian and decentralist ideas. Moreover, some potentially testable hypotheses have been posed.

REFERENCES

William J. Baumol, *Business Behavior Value and Growth,* New York 1959.

E. M. Drandrakis and Edmund S. Phelps, "A Model of Induced Invention, Growth, and Distribution," *Econ. J.*, June 1966, *71*.

William Fellner, Measures of Technical Progress in the Light of Recent Growth Theories," *Amer. Econ. Rev.*, Dec. 1967, *57*.

John Kenneth Galbraith, *The New Industrial State,* Boston 1971.

Charles Kennedy, "Induced Bias in Innovation and the Theory of Distribution," *Econ. J.*, Sept. 1964, *74*, 541–47.

Roger A. McCain, "Induced Technical Progress and the Price of Capital Goods," *Econ. J.*, Sept. 1972, *82*.

William Nordhaus, "Some Skeptical Thoughts on the Theory of Induced Innovation," *Quart. J. Econ*., May 1973, *87*, 208–19.

E. Schumaker, *Small is Beautiful,* New York 1973.

SOME ASPECTS OF INCOME DISTRIBUTION

The Effects of the Rural Income Maintenance Experiment on the School Performance of Children

By Rebecca A. Maynard*

A comprehensive evaluation of any welfare reform proposal must consider both long-run and short-run costs and benefits. One short-run benefit of a negative income tax program *(NIT)* is likely to be improvements in the school performances of the children of participants. These improvements may occur for several reasons: health may improve as a result of increased consumption of nutritious food and health care; learning aids such as books and magazines may become more readily available; and parents may spend more time with their children, participating with them in learning-related activities. Therefore, a long-term benefit of an *NIT* may well be the higher earnings that these children will eventually receive as a result of such improvements

This paper summarizes the findings of an analysis of the effects of the Rural Income Maintenance Experiment on four measures of school performance—attendance, comportment grades, academic grades and standardized achievement test scores.

I. The Model

The analysis of the relationship between participation in a negative income tax program and school performance is based on a theoretical model which assumes that learning is composed of two distinct but interrelated components: learning that occurs as a result of environmental exposure rather than as a consequence of any behavioral choice and learning that is directly attributable to the child's choice of activities. Changes in the home and/or school environments may affect both components of learning. If the general intellectual quality of the environment improves, the child's nonchoice learning will increase. Also, changes in the environment will affect the child's allocation of time among activities that differ in terms of their learning components.

Notationally, the determinants of learning can be expressed as follows:

$$L_t = f(L_{t-1}, g, A_t, E_t)$$

where

L_t = learning in period t

L_{t-1} = the knowledge stock in period t

g = genetic endowment

A_t = activities selections in period t, and

E_t = environmental conditions in period t.

The introduction of a negative income tax program may affect learning directly and indirectly through its effects on the child's environment and on the child's activity selections. The relationships between the program parameters and the variables in the learning function are discussed briefly below.[1]

*Senior Economist, Mathematica Policy Research, Princeton, N.J. This research was supported by the Institute for Research on Poverty and Mathematica Policy Research. I am indebted to Richard Murnane and Harold Watts for their comments and suggestions

[1]Maynard presents a more rigorous exposition of this theoretical model

The parameters of a negative income tax program include the guaranteed annual income, G, and the income tax rate, t. If total earnings and other income falls below the critical level determined by G and t, the family receives an income subsidy which is a function of the program parameters and total earned and other unearned income. Since payments are made to the parents, the program's impact on school performance depends indirectly on the induced modifications in the parents' behavior.

For families whose incomes fall below the break-even level, the primary effects of the program are expected to be 1) an increase in total family income and 2) a reduction in parents' labor force activity. These effects can be broken down into a price effect of the tax, an income effect of the (compensated) tax, and an income effect of the guarantee.

The price effect of the tax on income lowers the effective wage rate, thereby inducing a reduction in both labor force activity and in income The income effect of the tax tends to offset the price effect by inducing an increase in these factors. The usual assumption is that the net effect of the imposition of or increase in the tax will be a simultaneous decrease in labor force hours and in income

These two effects are competing in terms of their effects on school performance. It is assumed that parental time and income expenditures that complement education objectives are normal goods and superior to expenditures that compete with education objectives Thus, parents will spend the additional nonlabor-market time 1) interacting generally with their children, 2) in ways that improve the child's learning efficiency, 3) in ways that increase the desirability of learning activities, and 4) substituting for the child's work-type activities. The first of these effects will lead to greater amounts of nonchoice learning, while each of the last three effects will alter the child's time allocation between activities of greater and lesser learning intensities.

The tax-induced reduction in income is assumed to have a negative influence on the environment, with the result that less of the child's time will be devoted to learning-intensive activities and less nonchoice learning will occur. It is uncertain whether the favorable price effect of the tax will dominate this negative income effect.

However, for families initially below break-even, it is assumed that the tax-induced decrease in income is more than offset by the guaranteed annual income of an *NIT* program. To the extent that parents allocate this income to the consumption of goods that increase the quality of the environment and that encourage the consumption of learning-intensive activities, school performance will improve. For example, parents may purchase goods such as increased non-labor-market time, better quality housing, more nutritious foods, health care, and books. These goods may improve the child's learning efficiency, and they may permit and encourage the consumption of greater amounts of learning-intensive activities. Also, constraints on school attendance and continuation may be alleviated

In summary, imposition of a negative income tax is expected to lead to improved school performance through its effects on the parents' time and income allocations. Their time reallocations may result in an increase in 1) knowledge diffusion from the parent or other environmental aspects to the child, 2) the release of time constraints on the child, or 3) increases in the relative desirability of learning intensive activities. The Rural Income Maintenance Experiment collected data to investigate whether or not such positive influences on performance are observed when a negative income tax program is implemented.

II. The Sample and Data

The sample of children used for the analysis is a subset of the children whose families participated in the experiment. It includes all children 1) who were in grades 2 through 12 at the time of the most recent observations on any school performance measure, 2) who did not change their household of residence during the

experiment, and 3) whose family background data are relatively complete. Further, the analysis of any particular dimension of school performance includes only children for whom pre- and postenrollment performance data are available. Altogether, 847 children were included in some portion of the analysis.

The sample children are not representative of the student population in their respective regions —Duplin County, North Carolina and Pocohantas and Calhoun Counties, Iowa Furthermore, they are not representative of the nation's population of school children but of an intellectually impoverished population For example, low family income, low parental education levels and large family sizes typify both this sample and students who have a relatively high risk of school failure. Table 1 shows selected characteristics of the Iowa and North Carolina experiment samples, the Iowa and North Carolina regions, and the nation. The differences between the analysis samples and the larger populations affect the generalizability, but not the validity, of the results.

The data base used in the analysis includes extensive information on the school performance of children and the resource supply characteristics of their schools and (for North Carolina grade school children) classrooms. In combination with information concerning each child's home environment, these data have permitted a detailed investigation of the determinants of school performance and the calculation of relatively precise estimates of the effect of the negative income tax program on each of four measures of performance—absenteeism, comportment grades, academic grades, and standardized achievement test scores.

TABLE 1—CHARACTERISTICS OF THE RURAL INCOME MAINTENANCE EXPERIMENT'S SAMPLE, THE SITES AND THE UNITED STATES

Characteristic	North Carolina		Iowa		United States
	Experiment Sample	Duplin County[a]	Experiment Sample	Pocahantas County[a]	
Total Family Income ($)	3,645.00	6,085 00	3,997 00	9,591 00	9,433 0[c]
% Families with Income Below Poverty Level	62 10	39 90	36 80	9 70	9.7[c]
Education of Head (years)	7 60	8 88	10 90	11 50	12 3[c]
% Female Heads	12 70	18 90	13 20	6 80	10 8[b]
% Farmers	29 90	28 10	49 80	83 30	5 2[b]
Family Size	6 50	4 90	6 10	5 10	3 6[b]
Rooms/Person	0 90	0 94	0 92	0 78	N A
% Black	67 90	43 70	0 00	0 00	11 2[b]

Note. N A = not available.
[a]*Source.* the screening interviews for the Rural Income Maintenance Experiment
[b]*Source: Statistical Abstract of the United States*, U S Bureau of the Census, 1970
[c]*Source Statistical Abstract of the United States*, U S Bureau of the Census, 1971

III. The Results

The general form of the model used to estimate the effects of the negative income tax program on school performance is:

$$L_t = \alpha + \beta_1 L_{t-1} + \sum_{i=2}^{k} \beta_i \ln E_i + \sum_{j=k+1}^{n} \beta_j E_j + \beta_{n+1} T + \mu$$

TABLE 2—ESTIMATED DIFFERENTIALS IN SCHOOL PERFORMANCE ADJUSTED MEANS FOR EXPERIMENTAL AND CONTROL GROUPS[a]

Measure of School Performance	North Carolina				Iowa			
	Experimentals	Controls	Differential $(E-C)$	Percent Differential $100x(E-C)/\|C\|$	Experimentals	Controls	Differential $(E-C)$	Percent Differential $100x(E-C)/\|C\|$
Days Absent from School —Grades 2 through 8 (annual)	9.13	13.13	−4.00**	−30.46**	7.54	9.39	−1.85	−19.70
Days Absent from School —Grades 9 through 12 (annual)	7.85	7.61	0.24	3.15	6.80	8.19	−1.39	−16.97
Comportment Grade Point Average[b]	233.42	218.81	14.61**	6.68**	229.37	230.30	−0.93	−0.40
Academic Grade Point Average—Grades 2 through 8	225.92	212.79	13.13*	6.17*	249.90	261.95	−12.05	−4.60
Academic Grade Point Average—Grades 9 through 12	203.45	195.84	8.41	4.29	244.25	255.97	−11.72	−4.58
Deviation from Expected Grade Equivalent Score on Standardized Achievement Test[b]	−14.36	−17.71	3.35**	18.92**	−1.22	1.38	−2.60	−188.41
Percentile Score on Standardized Achievement Test[b]	25.83	25.43	0.40	1.57	44.73	52.47	−7.74	−14.75

*Statistically significant at the ten percent level

**Statistically significant at the five percent level

[a]Control variables included in the regressions are pre-enrollment measures of (1) the dependent variable, (2) personal characteristics such as ethnicity, sex, *grade level*, health status, and sibling position, (3) family background characteristics such as *parental education*, farm status, and *family size* and (4) home environment factors such as *parents' time allocation*, *income*, *assets*, and *nutrition*; and during program school environment characteristics such as school/class enrollment, performance on standardized tests, *ethnic composition*, and teachers' characteristics The logarithms of variables in italics were included in the analysis.

[b]Comportment grades and standardized achievement test scores are avilable only for children in grades 2 through 8.

where

L_t and L_{t-1} = the post- and pre-experiment performance measures, respectively

E = environmental conditions prior to enrollment in the experiment and school environmental conditions during the period of time between the pre- and postenrollment measures of performance

T = a treatment status variable, and

μ = a random error term

All but the absenteeism equations were estimated using ordinary least squares. Since a high degree of intrafamily correlation among the error terms in the absenteeism equations was observed, those equations were estimated using an error-components model that includes a family-specific error term.

The main findings with respect to the program response are summarized in Table 2, which presents adjusted mean estimates of school performance for children from experimental families and for children from control families.[2] The largest and most significant responses were found for the sample of 2nd through 8th grade students from North Carolina families These children exhibited statistically significant responses for all four measures of school performance. Positive experimental responses in terms of relative differences in the adjusted mean values of outcome measures include a 30.5 percent reduction in absenteeism, a 6.7 percent increase in comportment grade point average, a 6.2 percent increase in academic grade point average, and an 18.9 percent improvement in the deviation between achievement test scores and expected grade equivalent scores. The percentile scores on standardized achievement tests also rose for 2nd through 8th graders in the North Carolina experimental group relative to children in the control group, but the adjusted mean differential is relatively small and statistically insignificant.

The older sample of children from North Carolina families did not exhibit any significant experimental responses. The adjusted mean value for academic grades was 4.3 percent higher for children from experimental families, but this finding is not statistically significant. The differential for 9th through 12 graders in terms of number of days absent per year is very small and not statistically different from zero.

While the above estimates of experimental responses for children from North Carolina families support the general hypothesis that a negative income tax program will result in improved school performances, the findings for the Iowa samples provide little or no support for this hypothesis None of the adjusted mean differences in school performance between the experimental and control groups in Iowa is statistically significant. Children from experimental families did reduce their absenteeism, but there was essentially no difference in comportment grades between experimentals and controls. However, the most puzzling result is that both academic grade point averages and achievement test performances tended to be lower, other things being equal, for children whose families were enrolled in experimental negative income tax program. At least part of this anomalous finding can be explained by the presence of differences in sample characteristics[3] and the poorer quality of the Iowa school environment data.[4]

[2]The responses are presented in terms of adjusted means so that the results can be interpreted as applicable to experimental and control groups that have identical compositions, equivalent to the mean values of the control variables

[3]In general, students in the Iowa sample compared with those in the North Carolina sample were better performers prior to their enrollment in the experiment While, within each sample, the relationship between the pre- and postexperiment measures of performance seems to be linear, it may be that nonlinearities do exist if the entire range of performances is considered (for example, if data from the two regions are pooled)

[4]Student-specific data on classroom and teacher characteristics are available for the North Carolina sample, but only school-specific demographic data such as per pupil expenditures and enrollment are available for the Iowa sample

In general, the program responses tend to be larger and more significant for children in the lower grade levels whose behaviors are easier to modify. Also, for children in all grade levels, the program responses are larger for those performance measures, such as absenteeism and comportment grades, that more closely embody current behavior and are less dependent on the effects of past behavior through the stock of knowledge. Investigations of whether or not the program response varied depending on the expected level of benefit or preprogram characteristics of the individual or his/her environment revealed only one consistent finding: as the expected payment level increased, absenteeism decreased still further. Some significant relationships between preprogram factors and the treatment response were identified for selected outcome measures and samples. However, these findings do not appear to be generalizable.

IV. Conclusion

The results of this study suggest that the introduction of a national *NIT* program may lead to overall improvements in school performance and increases in the levels of educational attainment. Participation in the experimental group favorably influenced the performance of the 2nd through 8th graders in the North Carolina sample. These children demonstrated significant improvements in attendance, comportment grades, academic grades and standardized achievement test scores. The most noteworthy of these effects is the 30 percent reduction in their absences. However, neither the high school-age children in North Carolina nor the children in Iowa demonstrated any significant change in their performance as a result of participation in the program.

The findings are not without anomalies which, given the potential magnitude involved, clearly establish the need for further research in this area. For example, some puzzling differences in both the magnitude and direction of responses across sites have not been adequately explained. Before more conclusive policy inferences can be drawn, additional research is required both on the general determinants of school performance and, more specifically, on the causal effects of a negative income tax

REFERENCES

Rebecca A. Maynard, "A Theoretical and Empirical Investigation of the Effects of an Income Maintenance Program on School Performance," Mathematica Policy Research Working Paper, No. C-8, 1976

U.S. Bureau of the Census, Department of Commerce, *Statistical Abstract of the United States*, Washington, D.C., U.S. Government Printing Office 1970, 1971

University of Wisconsin, Institute for Research on Poverty, Screening Interviews, The Rural Income Maintenance Experiment, 1969.

Sons of Immigrants: Are They at an Earnings Disadvantage?

By BARRY R. CHISWICK*

In 1970, 9.6 million persons in the United States, or 4.6 percent of the population, were foreign born. Another 24 million persons, or 11.5 percent of the population, were of foreign parentage, that is, either one or both parents were foreign born. The earnings and labor market behavior of the foreign stock (foreign born and foreign parentage) have not been the subject of much systematic research despite the rise in public interest in ethnicity and discrimination. This paper, which focuses on the foreign parentage, is drawn from a larger study of the earnings of the foreign stock which is intended to remedy this situation (see Chiswick).

This paper examines the effect of foreign parentage on the earnings of native born white men age 25 to 64 who worked in 1969. It is restricted to whites as they comprise 97 percent of the persons of foreign parentage and to men because the problems of estimating labor market experience for women require that they be dealt with separately. In addition, persons born in Puerto Rico or an outlying area of the United States are excluded from the data.

I. Variables and Hypotheses

Among native born white men age 25 to 64, 19 percent had at least one foreign born parent. These men had significantly higher annual earnings (12 percent) in 1969 than men with native born parents (Table 1). Although there is virtually no difference in the number of years of schooling or weeks worked during the year between the two groups, the foreign parentage tend to be older and disproportionately concentrated in urban areas and outside the South where wages are generally higher. Multivariate analysis is needed to disentangle the effects on earnings of these explanatory variables.

Having foreign born parents may have a direct effect on earnings, holding constant easily measured dimensions of labor quality. There may be beneficial effects acquired from one's parents (through inheritance or environment) due to the selectivity bias in migration, that is, the tendency for migrants to be disproportionately high ability or highly motivated persons. There may, however, be disadvantages arising from being raised in a home less familiar with the language, customs, and institutions of the United States. In addition, there may be discrimination against second generation Americans in wages, employment, or union membership. A priori the net effect of these factors is uncertain.

The effect of having a foreign parent may depend on whether it was the father, the mother or both who were foreign born. There are several hypotheses that would imply a differential effect of parents' nativity. Foreign born fathers are more likely to have migrated for their own economic reasons, are therefore not a random sample of the men from their country of origin and are likely to be of higher innate ability or motivation than men native to the country of destination. Then, if labor market ability and aspirations among native born men are influenced by their fathers, having a foreign born father would tend to be associated with higher earnings. If their knowledge of the country's culture, language, etc., is more heavily influenced by their mothers, however, having a foreign born mother would tend to decrease earning potential. Another hypothesis that suggests an "effect" of mother's nativity given that the father is foreign born is that the most financially successful of foreign born men may marry native born women, and there is a positive association of characteristics that generate high earnings in fathers and sons.

*Senior Staff Economist, Council of Economic Advisers. The views expressed in this paper are solely those of the author and are not to be attributed to the CEA. James Moser's research assistance is appreciated. Comments received from Carmel U. Chiswick have, as usual, been most helpful.

TABLE 1—CHARACTERISTICS OF NATIVE BORN WHITE MEN, AGE 25 TO 64, BY NATIVITY OF PARENTS, 1970

Variable	All		Both Parents Native Born		One or Both Parents Foreign Born	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
1) Earnings ($)	9,653.53	7,600.69	9,441.93	7,417.57	10,567.98	8,284.73
2) Log of Earnings (in thousands of $)	4.31	0.86	4.29	0.86	4.41	0.84
3) Education	11.91	3.42	11.92	3.43	11.85	3.35
4) Age	42.72	11.09	41.71	11.08	47.08	10.06
5) Experience (age-education-5)	25.81	12.33	24.79	12.31	30.23	11.42
6) Weeks Worked	48.27	7.85	48.27	7.85	48.28	7.86
7) Log of Weeks Worked	3.85	0.29	3.85	0.29	3.85	0.29
8) Percent Rural	30.24	45.93	33.14	47.07	17.70	33.17
9) Percent South	29.24	45.49	33.60	47.24	10.37	30.49
10) Percent Not "Married, Spouse Present"	14.44	35.15	14.55	35.26	13.96	34.66
11) Percent Father Foreign Born	16.24	36.88	N.A.	N.A.	86.41	34.27
12) Percent Mother Foreign Born	13.91	34.60	N.A.	N.A.	74.02	43.86
13) Percent Both Parents Foreign Born	11.36	31.73	N.A.	N.A.	60.43	48.90
14) Percent in which a Language Other than English was Spoken in the Home	22.69	41.89	12.73	33.33	65.76	47.46
Number of observations	33,878		27,512		6,366	

Note: N.A.—Not applicable

Source: 1970 Census of Population, 1/1,000 Sample, 15 percent questionnaire

II. The Analysis

The analysis uses the human capital earnings function as the point of departure. This function was originally introduced at the 1965 American Economic Association meeting (Gary Becker and Chiswick), but it has since been expanded (Jacob Mincer) and has become a widely used tool of analysis. The basic equation used here is a linear regression of the natural log of annual earnings (wages, salary and self-employment income expressed in thousands of dollars, *LnE*) on the exogenous variables:

1. *EDUC* — Years of schooling completed.
2. *T* — Labor market experience, measured as age-schooling-5.
3. *TSQR* — Experience squared.
4. *LNWW* — The natural log of weeks worked.
5. *RURALEQ*1 — Dichotomous variable equal to unity for a person living in a rural area, and zero otherwise.
6. *SOUTHEQ*1 — Dichotomous variable equal to unity in the 17 southern states, including the District of Columbia, and zero for other states.
7. *NOTMSP* — Dichotomous variable equal to zero for a person who is married, spouse present, and unity otherwise.
8. *PARFOR*, *MOFOR*, *FAFOR*, *BOPFOR* — Dichotomous variables equal to unity if either parent, the mother, the father, or both parents, respectively, are foreign born.
9. *NONENG* — Dichotomous variable equal to unity if a language other than, or in addition to,

English was spoken in the home when the person was a child.

Table 2 presents the multiple regression analyses using data from the 1/1,000 Sample of the 1970 Census of Population. Columns (1) and (2) present regression equations computed separately for the native and foreign parentage. The equations are remarkably similar. They have the same explanatory power and the differences in the regression coefficients are very small and not statistically significant. The difference in t-ratios is consistent with the different sample sizes. The one exception is the region variable. While living in a southern state has a significant depressing effect (11 percent) on earnings for men of native born parents, for men of foreign parentage the magnitude is smaller (5 percent) and hardly significant ($t = -1.65$).

TABLE 2—ANALYSIS OF EARNINGS OF NATIVE BORN WHITE MALES, 25 TO 64 YEARS OF AGE, BY NATIVITY OF PARENTS, 1970

(Dependent variable: natural logarithm of earnings expressed in thousands of dollars)

	Regressions						
	Parents Native Born	At Least One Parent Foreign Born			Parents Native or Foreign Born		
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EDUC	0.06948	0.07220	0.07235	0.07132	0.07017	0.06988	0.06967
	(47.15)	(23.18)	(23.22)	(22.66)	(52.8)	(52.54)	(52.14)
T	0.03216	0.02511	0.02554	0.02585	0.03108	0.03062	0.03065
	(20.69)	(7.02)	(7.06)	(7.14)	(22.09)	(21.71)	(21.72)
TSQR	–0.00054	0.00038	–0.00038	–0.00039	–0.00051	–0.00051	–0.0051
	(–18.89)	(–6.22)	(–6.24)	(–6.37)	(–19.85)	(–19.74)	(–19.75)
LNWW	1.13749	1.14221	1.14307	1.14363	1.13902	1.13880	1.13902
	(74.66)	(36.96)	(36.98)	(37.01)	(83.31)	(83.32)	(83.33)
RURALEQ1	–0.18103	–0.16236	–0.16349	–0.16445	–0.18357	–0.17881	–0.17959
	(–19.22)	(–7.01)	(–7.03)	(–7.07)	(–21.22)	(–20.54)	(–20.60)
SOUTHEQ1	–0.10989	–0.04768	–0.04828	–0.04694	–0.11280	–0.10514	–0.10624
	(–11.85)	(–1.65)	(–1.67)	(–1.62)	(–13.05)	(–11.95)	(–12.05)
NOTMSP	–0.31711	0.32357	–0.32405	–0.32496	–0.31888	–0.31885	–0.31821
	(–25.30)	(–12.50)	(–12.52)	(–12.56)	(–28.22)	(–28.22)	(–28.14)
PARFOR	(a)	(a)	(a)	(a)	(a)	(a)	(a)
						(4.67)	
FAFOR	(a)	(a)	0.00902	0.02726	(a)	(a)	0.07688
			(0.33)	(0.97)			
MOFOR	(a)	(a)	–0.03270	–0.01823	(a)	(a)	0.03735
			(–1.53)	(–0.82)			
BOPFOR	(a)	(a)	(a)	(a)	(a)	(a)	–0.05873
							(–1.80)
NONENG	(a)	(a)	(a)	–0.04594	(a)	(a)	–0.01955
				(–2.28)			(–1.78)
Constant	–1.15699	–1.12442	–1.12213	–1.10936	–1.15892	–1.15770	–1.15346
Number of observations	27,512	6,366	6,366	6,366	33,878	33,878	33,878
R	0.55495	0.55103	0.55133	0.55184	0.55555	0.55595	0.55605
R^2	0.30797	0.30364	0.30396	0.30453	0.30863	0.30908	0.30919
Standard error	0.71947	0.70202	0.70196	0.71073	0.71652	0.71630	0.71628

Note: (a) Variable not entered. t-ratios in parentheses.
Source: 1970 Census of Population, 1/1,000 Sample, 15 percent questionnaire

If the effect of region is viewed as a compensating differential for the nonmoney aspects of living in an area, the different effect of southern residence on earnings implies a weaker preference for living in the South among the children of the foreign born, and presumably also the foreign born themselves. The smaller effect on earnings of residing in the South for the foreign born may also reflect a difference in real incomes. Migrants to the United States are more likely to settle in the region with a higher real income and greater employment opportunities. Historic roots may have retarded the regional adjustment of persons whose ancestors have lived in the South for several generations. That is, the substantial outmigration from the South may not have been sufficient to equalize earnings. Either hypothesis is consistent with the data on regional distribution by parents' nativity; 10 percent of the foreign parentage and 34 percent of the native parentage white men live in the South.

Pooled native and foreign parentage regressions are reported in columns (5) to (7). Other things the same, earnings are 5 percent higher for persons with one or both parents foreign born. Compared to men with native born parents, earnings are higher by 7.7 percent if only the father is foreign born, 5.6 percent if both parents are foreign born, and 3.7 percent if only the mother is foreign born. These results suggest that whatever disadvantages persons of foreign parentage have (e.g., less information about the United States, discrimination by others) appear to be overcome by other factors, particularly if the father is foreign born.

Earnings tend to be lower if a language other than, or in addition to, English was spoken in the home when the person was a child. The effect is a weakly significant 2 percent in the analysis of all native born persons, but it is larger, 4.6 percent, and more significant, if one or both parents are foreign born.

A decomposition analysis can be used to explain why, on average, foreign parentage men have earnings that are 12 percent higher than native parentage men. Evaluated at the mean, 2.5 percentage points of the difference arises because the foreign parentage are more likely to live in an urban area. They also have higher earnings because they are less likely to live in the South (2.6 percentage points using the native parentage coefficient), and because living in the South has a weaker depressing effect on their earnings (.6 percentage point, or 1.1 percentage points if the foreign parentage coefficient is set equal to zero). These two variables explain about one-half of the higher earnings of the foreign parentage. Using native parentage coefficients, the greater labor market experience of the foreign parentage would tend to raise their earnings by 2.5 percent, but this may be offset by the flatter experience earnings profile (4.9 percent), although the slope coefficients do not differ significantly. The schooling, weeks worked and marital status variables do not provide an explanation of earnings differences.

Is it possible that earnings vary systematically with the country of origin of the parents? Parents' country of birth is coded by the Bureau of the Census as the father's if he was foreign born, and the mother's if the father was native born. The frequency distribution of parents' country of birth for the group under study is:

TABLE 3

Country	Percent	Country	Percent
British Isles	11.6	Canada, Australia,	
Western Europe	23.9	New Zealand	7.8
Southern Europe	19.5	Mexico	5.1
Central Europe	16.3	Other Latin America	0.5
Russia	9.6	Asia and Africa	1.6
Balkans	4.1	Total	100.0

Source: 1970 Census of Population, 1/1,000 Sample, 15 percent questionnaire

Using the British Isles as the benchmark group, parents' country of origin dummy variables were added to the foreign parentage regression equation. The t-ratios were less than unity for all of the major sources of immigration, except Mexico. *Ceteris paribus*, earnings are lower by 18 percent ($t = -4.1$) for those with a Mexican

born parent. However, native parentage Mexican-Americans have lower earnings than other white men of native born parents, *ceteris paribus.* (When a dichotomous variable for Spanish surname men living in the Southwest is added to the native parentage regression, it has a slope coefficient of $-.268$ ($t = -6.06$). This does not differ significantly from the Mexican origin effect in the foreign parentage analysis.) The lower earnings of foreign parentage Mexican-Americans reflect a characteristic of an ethnic group, rather than a problem peculiar to second generation status.

III. Summary and Conclusions

The analysis suggests that second generation white male Americans differ very little from white males of native born parents, if anything, having a slight earnings advantage (5 percent) when other things are held constant. The annual and weekly earnings of the foreign parentage are higher by almost 12 percent. The foreign parentage, however, are less likely to live in a rural area or in the South where earnings are lower, and this explains about half of the observed difference in earnings. Although on average they are five years older, the foreign parentage have about the same level of schooling and work the same number of weeks in the year. With the exception of the South-non-South variable, the partial effects of the explanatory variables are the same for both groups.

Earnings are higher by about 8 percent if the father is foreign born and the mother is native born, in comparison with those of native parentage. The advantage from having foreign parents is about 6 percent if both are foreign born and 4 percent if only the mother is foreign born. Other things the same, earnings are lower (by about 2 to 5 percent) if a language other than English was spoken in the home when the person was a child. Persons of Mexican parentage do have significantly lower earnings, *ceteris paribus,* but this appears to be a problem shared in common with Mexican-Americans whose families have lived in the United States for three or more generations.

If there is discrimination against second generation Americans it appears to be overcome by other factors. These other factors may include innate ability, or knowledge, skills or motivation acquired in the home. Some of the factors that are associated with the higher rate of migration of the more able or more highly motivated may be passed on through inheritance or environment to the migrants' native born children. It would appear, however, that there is a regression to the mean in these characteristics from one generation to the next

REFERENCES

Gary S. Becker and Barry R. Chiswick, "Education and the Distribution of Earnings," *Amer. Econ. Rev. Proc.*, May 1966, *66*, 358–69.

Barry R. Chiswick, "The Effect of Americanization on Earnings," July 1976, mimeo.

Jacob Mincer, *Schooling, Experience and Earnings*, New York 1974.

U.S. Bureau of the Census, *1970 Census of Population*, One-in-a-Thousand Sample, 15 percent questionnaire.

Short-run Housing Responses to Changes in Income

By ELIZABETH A. ROISTACHER*

Income maintenance and housing allowance plans, both currently being evaluated as potential federal programs for low-income families, provide households with supplements to income which increase the demand for housing. While income maintenance schemes allow the consumer complete freedom of choice with respect to consumption, housing allowance plans may constrain choice in a variety of ways.[1] The response of housing demand to changes in income is a significant factor in assessing the relative merits of these two programs

The research presented here is concerned with the impact of increased income on the annual housing expenditures of households who have moved. The data employed are drawn from the University of Michigan panel study of income dynamics, which follows a sample of 5,000 households over a period of years. These panel data allow us to link each household's change in income over time with changes in its demand for housing. Most previous studies of housing demand rely on cross-sectional estimates which are long run: they implicitly assume a household would respond to a different income level just as would a household who may have been at that level for a long enough period of time to acquire a different set of tastes. To the extent that households move from one short-run adjustment to another because of repeated changes in economic circumstances, rather than ever actually achieving the implied long-run equilibrium, short-run elasticities may be more revelant for evaluating housing policy.

I. A Model of Housing Demand Over Time

Increases in a household's income may affect housing demand in a number of ways. First, additional income could induce upgrading of its current unit through higher maintenance or through additions or modifications to the structure Upgrading is most likely to occur among homeowners. Alternatively, a household may make a major adjustment in housing consumption, possibly including a change in tenure, by moving to another dwelling unit. Because of the high transactions costs associated with moving —search costs and actual moving costs—, households are likely to move relatively infrequently, so that actual housing consumption may lag behind its desired level. It seems reasonable to assume, however, that households who do move will attempt to adjust fully to their new economic circumstances, or even "overadjust" in anticipation of future changes, because of the high transactions costs. Therefore, households who move are assumed to make at least a complete short-run adjustment, given their current tastes, although in the long run their tastes may change.[2]

*Assistant Professor of Economics, Queens College, *CUNY* This research was supported by a Faculty Research Award Craig Winderman ably performed the computer work and provided invaluable research assistance George Borjas, Greg Duncan, and William DuMouchel provided helpful advice

[1]The Housing Allowance Demand Experiment sponsored by the U S. Department of Housing and Urban Development employs both "percent of rent" and "housing gap" formulae to determine payment In some cases "minimum standards" conditions are placed on the payments See Stephen Mayo, pp A–6 to A–8.

[2]Less than full adjustment, however, would be likely if a household is moving between cities An initial move to a less than optimal unit may be followed by a second move to the desired unit once the household has familiarized itself with its new housing market Some evidence of this pattern is found in Roistacher 1974, p. 48. Since our analysis eliminates inter-standard metropolitan statistical area *(SMSA)* movers, the likelihood of short-run partial adjustment is minimized

While we have estimated residential mobility and tenure change models using the data from the Michigan panel, we limit further discussion to our analysis of the change in annual housing expenditure for households who have moved.[3] Limitations in the data preclude analysis of the extent to which housing adjustment is made by upgrading a current unit rather than by moving.

The change in annual housing consumption for families who move is estimated as a function of changes over time in income, family size, and relative prices.[4] Let the demand for housing services at a point in time be specified as

$$(1) \qquad E_t = a_t Y_t^b N_t^c P_t^d \prod_k Z_k^{v_k}$$

or equivalently

$$(2) \qquad \ln E_t = \ln a_t + b \ln Y_t + c \ln N_t + d \ln P_t + \sum_k v_k \ln Z_k$$

where E is annual real expenditure on housing services, Y is permanent or normal income, N is family size, P is the relative price of housing, and the Z_k are demographic variables such as age, race, sex, and marital status of head, while b, c, d, and the v_k are time-invariant parameters, and a_t is a time effect.[5] Then the change in housing consumption from time $t - \tau$ to time t is equal to

$$(3) \qquad \ln(E_t/E_{t-\tau}) = a' + b \ln(Y_t/Y_{t-\tau}) + c \ln(N_t/N_{t-\tau}) + d \ln(P_t/P_{t-\tau})$$

where $a' = \ln(a_t/a_{t-\tau})$. $\ln(E_t/E_{t-\tau})$, $\ln(Y_t/Y_{t-\tau})$, and $\ln(P_t/P_{t-\tau})$ respectively approximate the percentage change in housing expenditure, income, and prices, and the coefficients b and d are interpreted as income and price elasticities of expenditure on housing.[6]

Equation (3) assumes that households are in equilibrium at the two points in time over which the income elasticity is measured. However, since our empirical analysis observes households who may be in disequilibrium prior to the move, we include in our estimated model two variables which control for this possible disequilibrium: the first measures the time since the household's previous move and the second is an interaction between this variable and the percentage change in income.

II. Data Base and Definitions of Key Variables

The data used for the analysis are taken from the first six years of the Michigan panel study (1968–73) and are restricted to those families who moved within one of the 24 largest *SMSA*s from 1969 to 1972, the period over which changes in income and housing expenditures are examined.[7] Approximately 550 households fall into the final sample.

[3]These results are available in a more extensive paper (Roistacher 1976). Income changes as well as levels were found to be significant in predicting residential mobility as well as in predicting the switch from renting to owning for families who move

[4]Since we deflate annual housing expenditure by the change in the price of housing, the change in annual expenditure measures the change in annual consumption of housing services.

[5]The time-invariance of b and c is supported by empirical results. Cross-sectional regressions for 1969 and 1972 indicate that the elasticities of income, family size, and age, are stable over the two time periods The Z_k are assumed to be time-invariant because the analysis is restricted to households with the same head over the period

[6]The expression $\ln(1 + x)$ has the Taylor's expansion

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^{n-1}\frac{x^n}{n} + \ldots$$

so that if x is small $\ln(1 + x)$ is approximated by x. (See R G D Allen, p 456). The expression $\ln(E_t/E_{t-\tau})$ may be rewritten

$$\ln[1 + (E_t - E_{t-\tau})/E_{t-\tau}]$$

which is of the general form $(1 + x)$, as are the income and price terms While the approximation of $\ln(1 + x)$ is better for small values of x, the concept of elasticity itself breaks down for large values of x. When percentage changes are large, one must resort to a midpoint formulation for which $\ln(1 + x)$ is an excellent approximation and a perfectly reasonable alternative

[7]By restricting the analysis to households within each of the largest *SMSA*s, we can employ *SMSA*-specific deflators and relative price variables as well as control for differences in supply and tastes associated with metropolitan markets

Annual housing expenditure for homeowners is calculated as 6 percent of reported house value plus annual property taxes and utilities.[8] The annual housing expenditure variable for renters is equal to annual rent plus annual utilities. While our measure of annual housing expenditures is likely to overstate true annual housing outlays for owners as house value rises, the error is not correlated with the percentage change in income; therefore, our estimated income elasticity is not biased by this error.[9]

Changes in income are calculated as the change in the average income from the first three years (1967–69) to the last three years (1970–72). This gives us a measure of the change in normal or permanent income.[10] (Each year's spring interview reports the previous year's income and current house value or monthly rent. (See Survey Research Center, *1*, pp. 283 and 307.) An imputed rent equal to 6 percent of net equity is included in the incomes of homeowners.

All income figures are adjusted to 1968 dollars using the Consumer Price Index *(CPI)* for the appropriate *SMSA*. Housing expenditures are deflated either by the renter component or the ownership component of the *CPI*. The change in the relative price of housing is calculated as the ratio of the change in either the renter or ownership component to the change in the overall *CPI*.[11]

Since we are particularly concerned with the responses of households who will be participants in government subsidy programs, we test for differences in behavior between low-income and all other households (hereafter referred to as high-income). A household is defined as low-income if it falls into the lowest decile of the population in *any* of the six years 1968–73 according to an income-to-needs criterion.[12]

III. The Estimated Model

For households who move we estimate the change in annual housing expenditure by ordinary least squares using the model

$$H_i = \sum_{j=0}^{7} \beta_j X_{ji} + \epsilon_i \tag{6}$$

where $H = \ln(E_t/E_{t-\tau})$ where E is annual expenditure on housing

$X_0 = 1$

$X_1 = \ln(Y_t/Y_{t-\tau})$ where Y is three-year average annual income

$X_2 = 1$ if $X_1 \leq 0$; zero otherwise

$X_3 = X_2 \cdot X_1$

$X_4 = \ln(N_t/N_{t-\tau})$ where N is family size

$X_5 = \ln(P_{j,t}/P_{j,t-\tau})$ where P is the relative price of housing

$X_6 =$ number of years since previous move

$X_7 = X_6 \cdot X_1$

with ϵ a normally distributed random error, i indicating the household, and j the *SMSA* in

[8] Leslie Kish and John B. Lansing have found no systematic bias in estimates of house value made by owner-occupants.

[9] The annual cost of housing falls as a percentage of house value for a number of reasons: economies in insurance, maintenance and utility costs, lower loan-to-value ratios and lower mortgage maturities as well as savings on property tax and interest costs from the increased tax benefits for higher income households. (See Frank de Leeuw.) For households switching from renting to owning the error in our dependent variable is approximately proportional to the measure of house cost, but since $\ln E_t$ is not strongly correlated with $\ln(Y_t/Y_{t-\tau})$, our income coefficient is not biased by this error (correlation coefficient = .13). (Initial owners who move are omitted from the analysis.)

[10] Regressions employing the change in three-year average annual income produced larger, more significant income coefficients than did regressions employing the change in current income or those which replaced current income with lagged income.

[11] The renter index is subject to "aging bias" so that it overstates true price increases. However, this bias has little impact over short periods of time. See Bureau of Labor Statistics, p. 4.

[12] This is the Orshansky-type poverty threshold which relies on a U.S. low-cost food plan with adjustments for economies of scale in food and non-food items. See Survey Research Center, 2, pp. 9–10. Because the Michigan panel oversamples lower income households, about 43 percent of the observations in our final sample fall into our definition of low income.

which the household lives. X_2 and X_3 allow tests for differences in slope and income elasticity for households whose real incomes decline; X_6 and X_7 adjust for households who may be in disequilibrium prior to the move, as discussed earlier.

Since elasticities may depend on the initial and final tenure of the household, we have stratified the households according to tenure change. A Chow test allows us to reject the hypothesis that households moving within the rental sector have the same coefficients as those switching from renting to owning. No equations are presented for households moving within the owner's sector or for those who switch from owning to renting; these results are statistically insignificant due to an insufficiency of cases. We further stratify the households moving within the rental sector into low- and high-income groups since a Chow test indicates that their coefficients are not the same. Unfortunately, there are too few low-income households switching sectors to allow for stratification of these households by income. Terms for income decreasers (X_2 and X_3) are insignificant for high- and low-income households moving within the rental sector, so regressions for these groups have been run with these terms omitted. The regression results are reported in Table 1. (For each group, regressions are reported with and without the terms for disequilibrium, X_6 and X_7.)

Equation (2) indicates that low-income renters have an income elasticity of .22 while equation (6) indicates that households who switch from renting to owning have a higher elasticity of .34. For both sets of households there is a marginally significant 2 percent increase in housing expenditures for every year since the previous move (see X_6).

TABLE 1—COEFFICIENTS FROM REGRESSIONS OF CHANGES IN ANNUAL HOUSING EXPENDITURES FOR SELECTED CATEGORIES OF MOVERS

Dependent Variable $\ln(E_t/E_{t-\tau})$

	Low-Income Households Renting in 1969 and 1972		High-Income Households Renting in 1969 and 1972		Households Renting in 1969 and Owning in 1972 (all incomes)	
Variable	*1*	2	3	4	5	6
$X_1 =$	27	22	07	04	.27	34
$\ln(Y_t/Y_{t-\tau})$	(3 43)	(2 39)	(83)	(40)	(1 99)	(1 78)
$X_2 = 1$					− 13	− 13
if $X_1 \leq 0$,					(−1.10)	(−1 11)
0 otherwise					− 86	− 86
$X_3 = X_1 \cdot X_2$					(3 34)	(−3 26)
$X_4 =$	05	06	31	.31	05	06
$\ln(N_t/N_{t-\tau})$	(61)	(73)	(4 52)	(4 47)	(3 47)	(51)
$X_5 =$	−1 08	−1 25	−1 32	−1 31	−2 87	−2 87
$\ln(P^h_{j,t}/P^h_{j,t-\tau})$	(86)	(− 99)	(−1 28)	(−1 26)	(−2 77)	(−2 76)
$X_6 =$		05		− 01		017
years since		(1 97)		(−.46)		(1 64)
previous move						
$X_7 = X_1 \cdot X_6$		02		01		− 016
		(98)		(50)		(− 82)
$X_0 = 1$	.01	−0.6	17	18	12	06
R-sq. (Adj.)	.061	.076	141	130	073	.078
Number of observations	173		137		158	

Source: Data from Panel Study of Income Dynamics, Institute for Social Research

Note t-ratios in parentheses

High-income households moving within the rental sector [equations (3) and (4)] appear to be completely insensitive to income changes but highly sensitive to changes in family size. These households may be saving for an eventual move to an owned home so that expenditures on housing in the short-run are not responsive to income changes.

For those households who switch sectors there is a significant and negative income elasticity when real income declines. [The elasticity for households with declining real incomes is $-.52 = .34 - .86$ from equation (6).] This could be evidence of a ratchet effect in housing consumption, of money illusion (the effect disappears if the regression is run without deflating), or of a feeling on the part of the household that the decline in income is only temporary. Since the effect is evident only among households switching sectors, however, it is difficult to generalize from this finding.

The apparent statistical significance of the relative price term in equations (5) and (6) is partially a result of spurious correlation with deflated housing expenditure, since the change in the price of owning is used as the expenditure deflator as well as for the construction of the relative price term. Since the price coefficient is biased downward (made more negative), it is dangerous to interpret this coefficient as evidence of a negative price elasticity of expenditure.[13]

IV. Comparisons with Other Estimates of Income Elasticity

We have found that low-income households moving within the rental sector and households switching from renting to owning have short-run income elasticities for housing expenditures of .22 and .34 respectively. However, given the standard errors of these estimates we cannot confidently reject the hypothesis that these coefficients are different. Higher income households moving within the rental sector appear to be insensitive to income changes, at least in the short run.

Short-run estimates of the income elasticity of housing demand have been made by Mayo, using data from the Housing Allowance Demand Experiment to estimate a stock adjustment model with housing consumption lagged one year. His short-run estimates for movers based on average income and on a regression-estimated permanent income range from .14 to .31, with corresponding long-run estimates ranging from .29 to .50. (Other estimates, however, lead him to put the long-run elasticity in the range of .4 to .6. See Mayo, pp. 84 and 99.) Using the Michigan data to estimate a stock adjustment model for movers, we find short-run elasticities ranging from .28 to .34 and long-run elasticities ranging from .40 to .49 when alternative one-, two-, and three-year lags in housing expenditures are specified (Roistacher 1976).

Most previous estimates of income elasticity are derived from static cross-sectional models in which inference about long-run elasticities are made across households within a single time period. The estimates range widely, from .4 to 2.0, with the higher estimates usually based on grouped data rather than on individual households. Two recent studies which used data from the Michigan panel to derive measures of permanent income have reported elasticities of .5 for renters and .7 for owners (Geoffrey Carliner) and .46 for both high and low-income renters when considered separately (Chester Fenton). A. Mitchell Polinsky has reviewed these and other cross-sectional studies and concludes that the misspecification of the relative price variable biased the results, which would otherwise cluster around .75. For lower income households, who move primarily within the rental sector, we can expect the long-run elasticity to be somewhat lower.

[13]Omission of the relative price term has virtually no effect on the other coefficients. In undeflated regressions the price term was not significant. Since the price variable takes on only 24 values, its variance is very low.

V. Conclusions

We have examined the short-run impact of increases in normal income on the annual housing consumption of households who move. We find that increases in housing consumption are relatively insensitive to changes in income: a 10 percent increase in income will produce only a 3 percent increase in housing consumption in the short run. Other research indicates elasticities are higher, but still inelastic.

The relative insensitivity of housing consumption to changes in income suggests that, if increased housing consumption is a specific governmental objective, then programs which lower the relative price of housing and/or direct consumption toward housing in some other way, in addition to supplementing income, would be preferable to cash transfers. Unfortunately, our results do not supply us with usable estimates of price elasticity, a parameter essential in assessing the relative effectiveness of alternative housing programs, although the Housing Allowance Demand Experiment may help to fill in this gap (see Mayo).

The demand of households participating in a full-fledged (nonexperimental) income maintenance program, however, may be more income-sensitive than indicated above, provided that the households involved perceive program subsidies as more permanent than the normal changes in income observed in our panel. Since findings from the New Jersey Graduated Work Incentive Program indicate some positive and significant experimental effects with respect to home purchase and expenditure on housing despite the short life of the experiment (see Judith Wooldridge), we can conclude that our estimates should be a lower bound of the potential short-run effects of an income maintenance program on the demand for housing.

REFERENCES

R.G.D. Allen, *Mathematical Analysis for Economists,* New York 1964.

Geoffrey Carliner, "Income Elasticity of Housing Demand," *Rev. Econ. Statist.*, Nov. 1973, *55*, 528–32.

Frank de Leeuw, "The Demand for Housing—A Review of the Cross-Section Evidence," *Rev. Econ. Statist.*, Feb. 1971, *53*, 1–10.

Chester Fenton, "The Permanent Income Hypothesis, Source of Income and the Demand for Rental Housing," Joint Center for Urban Studies, Cambridge, Mass., July 1974, 3–1 to 3–52.

Stephen K. Mayo, "Housing Expenditures and Quality," Part 1, Draft Report on Housing Expenditures Under a Percent of Rent Housing Allowance, Abt Associates, Inc., Cambridge, Mass., April 19, 1975.

A. Mitchell Polinsky, "The Demand for Housing: A Study in Specification and Grouping," Harvard Institute of Economic Research, disc pap no. 432, 1975.

Elizabeth A. Roistacher, "Residential Mobility," in James N. Morgan, ed., *Five Thousand American Families–Patterns of Economic Progress,* 2, Survey Research Center, University of Michigan, Ann Arbor 1974, 41–78.

———, "Shortrun Responses of Housing Demand to Changes in Income: An Analysis of Residential Mobility, Tenure Choice, and Housing Expenditures," Queens College, *CUNY*, 1976.

Judith Wooldridge, "Housing Consumption in the New Jersey-Pennsylvania Experiment," Final Report of the New Jersey Graduated Work Incentive Experiment, Institute for Research on Poverty, University of Wisconsin-Madison, and Mathematica, D3a–1–D3a–46.

Survey Research Center, *A Panel Study of Income Dynamics, 1* and *2*, Institute for Social Research, University of Michigan, Ann Arbor 1972.

U.S. Bureau of Labor Statistics, *The Consumer Price Index: History and Techniques,* Bulletin no. 1517, United States Dept. of Labor, Washington.

RADICAL ECONOMICS

Toward a Marxian Model of Economic Growth

By DAVID LAIBMAN*

In 1942, Joan Robinson urged us to the frontiers of our economic understanding by comparing Marxian and orthodox economics in these terms: "if there is any hope of progress in economics at all, it must be in using academic methods to solve the problems posed by Marx" (p. 95). In the spirit of that injunction, the present essay pursues the comparison on the limited terrain of models of economic growth. Its major finding is that an adequate conception of technical change as it occurs within capitalist production relations makes it possible to advance on both the Marxian and the growth-model fronts. With regard to the former, it is the key to overcoming seemingly insurmountable indeterminacies in the dynamics of the crucial variables; with regard to the latter, it brings into focus a capitalist-investor who is neither the neoclassical plaything of the sovereign consumer, nor the all-powerful *deus ex machina* of the post-Keynesians. The model suggested here will also transcend the limitations of the steady state, without losing coherence, although like the main prototypes of the neoclassical and post-Keynesian models it will skirt the complexities arising from a rigorous treatment of heterogeneous outputs and fixed capital

The Marxian ambiguities stem from Marx's unsubstantiated law of rising tendency of the "organic composition of capital"—the ratio of the value of the stock of physical capital to the flow of current labor cooperating with that stock in production For this ratio to rise, the accumulation of physical capital per unit of labor must exceed the rate at which rising productivity cheapens the elements of physical capital; in one-commodity macro terms, the physical capital-labor ratio, k, must grow more rapidly than the net productivity of labor, y Marx never supplied a rationale for this assertion, although he repeatedly affirms that the rising organic composition of capital is "but another expression for the rising productivity of labor" (On this see Maurice Dobb, Paul Sweezy and Laibman 1976) The "rate of exploitation," or ratio of the unpaid to the paid portions of the working day, is a second crucial variable which, despite much discussion, does not have clear determinants, although Marx's conception of the *relative* impoverishment of the working class implies that this variable should have a rising tendency. The well-known upshot is that the famous "law of the falling tendency of the rate of profit" is deprived of a consistent foundation. Nevertheless, the Marxian vision has unique synthetic power: choice of technique is inseparable from technical change, and both are aspects of the investment process, which in turn is a moment of the wider concept of *accumulation*, entailing not only quantitative growth but also the "extended reproduction" of antagonistic social relations in production, i.e , the accumulation of *capital* in the full Marxian sense of that term.

By contrast, the growth models of academic economics miss this vision, by either merging or rigidly separating its several aspects. In neoclassical growth, to illustrate, technical change

*Assistant Professor, Brooklyn College, City University of New York I wish to acknowledge the mathematical assistance of Ercument Ozizmir of Richmond College, and the computational assistance of Carl Lum of the Brooklyn College Computer Center. This paper is a report of a larger study in progress

is generally seen as autonomous and exterior to the economic process; choice of technique simply adapts the capital intensity of production to the rate of savings coming from sovereign consumers. Positive net investment is the passive result of an optimal allocation decision, like supplying consumers with anything else that has a positive demand price. The post-Keynesian models go to the opposite extreme, postulating full investor sovereignty and banishing the production function from their world altogether.[1] A Marxian growth model should avoid these opposite reductions of the capitalist, who will appear as a powerful subjective force in pursuit of identifiable goals but within a system that is a process without a subject, i.e., without totally unconstrained actors.

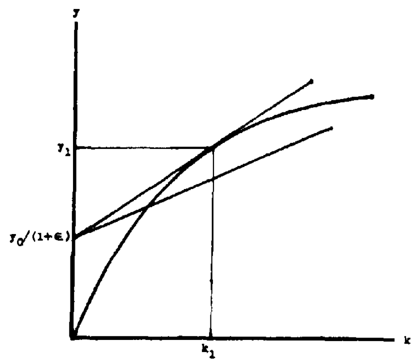


FIGURE 1

I

We begin by defining the rate of profit, as profits per head divided by the value of (physical) capital per head, noting that this implies wages paid out of revenue, and Marx's "variable capital" not subject to accumulation:

$$(1) \qquad r = y\left(\frac{\epsilon}{1+\epsilon}\right)\Big/k = y(1-2)/k$$

where ϵ is the rate of exploitation; $w = 1/(1+\epsilon)$ = the wage share of value added; y and k = physical net output and capital per head, respectively. We may note here that $1/y$ is the labor value of a unit of output; that the *value* of (physical) capital per unit of labor is therefore $k(1/y) = Q$, the organic composition of capital.

We also posit a *short-run* (in a sense to be defined below) technology constraint in the form of the familiar Cobb-Douglas relation.[2]

$$(2) \qquad y = c_0 e^{bt} k^{\alpha} \qquad 0 < \alpha < 1,\ b > 0$$

The key to technical change in the (capitalist) short run is that capitalists maximize the *transitional* profit rate, $\bar{r}$, defined as the rate of profit during the "transition period, when the use of machinery is a sort of monopoly, (in which) the profits are therefore exceptional . . .," accruing as they do to an innovator for whom the "social value of the article produced (is) above its individual value . . ." (Marx, pp. 443–44). If the individual capitalist purchases inputs from the outside, he will value them at the appropriate unit value $1/y_0$, where the zero subscript identifies the prevailing technique. The same will apply of course to the outputs. The maximand is therefore

$$(3) \qquad \bar{r} = \frac{y(1/y_0) - w}{k(1/y_0)} = \frac{y - wy_0}{k}$$

[1]The neoclassical literature may be well represented by Robert Solow (1956), Trevor Swan and James Meade. The leading post-Keynesian models have been propounded by Nicholas Kaldor and Robinson (1956, 1972). A seminal survey of the field is F. H. Hahn and R. C. O. Matthews.

[2]This function is defined for a short run, conceived as a period in which firms make decisions based on given prices; in the longer run, changes in valuation cause erratic or "perverse" movements in y and k, so that (2) does not have a "well-behaved" long-run counterpart, see Piero Sraffa, Geoffrey Harcourt. As a micro function, it should not be interpreted as a "pseudo" aggregate production function (Robinson 1956), but rather as a highly simplified expression of strictly engineering relationships; only the micro behavior which results need be aggregated. Contrary to appearances, linear homogeneity is not being assumed here; under nonconstant returns to scale the parameters of (2) would be functions of scale, but this would not affect the argument.

from which

$$(4)\qquad y = \bar{r}k + wy_0 = \bar{r}k + \left(\frac{1}{1+\epsilon}\right)y_0 .$$

This is a straight line in y-k space; since ϵ is exogenous to the firm, the intercept term $y_0/(1+\epsilon)$ is fixed. The firm will rotate (4) upward as far as possible, i.e., until it is tangent to the production function (see Figure 1). The tangency point thus identifies the target position of the firm, y_1 and k_1. At this point, $\bar{r}$ clearly equals the marginal product of capital, $\partial y/\partial k$. The marginal productivity principle, then, expresses the behavioral rule for the individual capitalist—maximize the *transitional* profit rate—and qualifies the neoclassical theory based on this rule as "bourgeois economics" in the precise sense of an economics whose limits are those which the capitalist class "does not transcend in practice."

Our task now is to find the parameters of y_1 and k_1, and then the growth rates of y, k and Q. First, the $\bar{r}_{\max}$ locus:

$$(4a)\qquad y_1 = y_0/(1+\epsilon) + \bar{r}_{\max}k_1 .$$

We will wish to assume that the production function shifts during the transitional period, although the shift will be small. The production function, before and after the shift, is written

$$(2a)\qquad y_0 = c_0e^{bt}k_0^{\alpha}$$

$$(2b)\qquad y_1 = c_0e^{bt}k_1^{\alpha}(1+bdt).$$

Lastly, we need the condition that the maximum transitional profit rate equal the slope of the (shifted) production function. From (2b) we find $\partial y_1/\partial k_1$ to get

$$(5)\qquad \bar{r}_{\max} = c_0\alpha e^{bt}k_1^{\alpha-1}(1+bdt) = y_1\alpha/k_1$$

Substituting into (4a), and rearranging:

$$(6)\qquad y_1 = y_0[(1-\alpha)(1+\epsilon)]^{-1}$$

from which

$$(7)\quad y^* = y_1/y_0 - 1 = [(1-\alpha)(1+\epsilon)]^{-1} - 1$$

where the asterisk denotes a growth rate. Putting (2a) and (2b) into (6) and solving for k_1:

$$(8)\qquad k_1 = k_0[(1+bdt)(1-\alpha)(1+\epsilon)]^{-1/\alpha}$$

from which

$$(9)\qquad k^* = k_1/k_0 - 1 = [(1+bdt)(1-\alpha)(1+\epsilon)]^{-1/\alpha} - 1$$

Finally, an expression for the growth rate of the organic composition of capital follows easily:

$$(10)\qquad Q^* = k^* - y^* = [(1+bdt)(1-\alpha)(1+\epsilon)]^{-1/\alpha} - [(1-\alpha)(1+\epsilon)]^{-1}$$

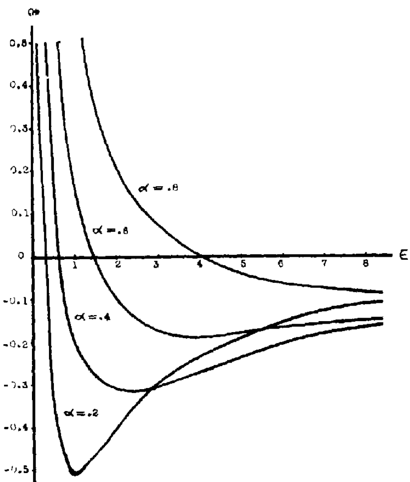


FIGURE 2

Figure 2 shows the shape of the Q^*—ϵ relation, with the "shift factor" $dt = 0$, for various values of α. The important feature here is the inverse relationship, with Q^* becoming zero at some (relatively high) level of ϵ. A "vertical integration" case, in which the

firm produces and values its inputs according to the novel technique, has also been analyzed; vertical integration shifts the curve outward, raising the critical value of ϵ at which growth of the organic composition of capital comes to a halt. In a Marxian world, α must be considered endogenous to the model, and the search for engineering practice which makes it possible to realize high transitional profits would lead this parameter to be high and perhaps rising. A high α, a small shift factor indicating intense competition, and progressing vertical integration all suggest that the critical ϵ is far above historically observed levels, and this amounts to a rather qualified "law of the rising organic composition of capital."

II

The Q^*—ϵ relation (10) is the—hitherto missing—link that will enable us to sketch the contours of a Marxian growth path for the capitalist economy. The other ingredients, savings and the growth of the labor force, present no new problems. As in all postclassical models, labor force growth is assumed constant and exogenous:

$$n = \text{constant} \tag{11}$$

The appropriate savings function is the well-known classical savings function, with zero savings out of wages and savings out of profits occurring at a high and constant rate a, determined by the interaction of "animal spirits" and the historically conditioned level of consumption out of profits. We therefore have

$$C^* = ar \tag{12}$$

where C is the *value* of the physical capital stock. From $Q = C/L$, where L is the labor force, we have

$$C^* = Q^* + n. \tag{13}$$

And finally, the basic identity of the Marxian system:

$$r = \frac{\epsilon}{Q(1+\epsilon)}. \tag{14}$$

Combining (12) and (13), we find the general form of what I will call the *consistent path* of the economy:

$$r = Q^*/a + n/a. \tag{15}$$

When $Q^* = 0$, this reduces to the Cambridge growth equation, $n = ar$. In general, $Q^* > 0$, and growth is off the steady state; hence the term "consistent path" in place of the usual "steady-state path."

By combining (15) with (10), we fully specify the consistent path (its derivation is diagrammed in Figure 3):

$$r = \frac{1}{a[(1+bdt)(1-\alpha)(1+\epsilon)]^{1/\alpha}} - \frac{1}{a(1-\alpha)(1+\epsilon)} + \frac{n}{a} \tag{16}$$

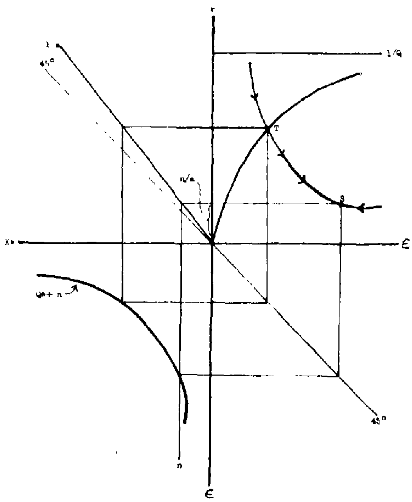


FIGURE 3

In the lower left quadrant of Figure 3, the growth factor $Q^* + n$ is plotted against ϵ, the

horizontal summation of (10 and (11). In the upper left, we have (12) and, linking the two left quadrants, (13), which ensures that the level of unemployment will be constant over time, since the growth rate of the demand for labor, $C^* - Q^*$, is equal to the growth rate of the labor supply n. The consistent-path relationship between r and ϵ is then derived in the upper right quadrant; it is the downward-sloping curve, whose curvature is determined by (10). The upward sloping curve in the same quadrant is (14), and it approaches the asymptote $1/Q$. The intersection of the two curves identifies a *consistent point T*, corresponding to some historically given level of Q. At T, $Q^* > 0$, and (14) is shifting downward. The initial impact of this is to lower r; with a fixed in the short period, C^* will fall, the growth rate of demand for labor $C^* - Q^*$ will be less than the growth rate of the labor supply, and unemployment will rise. For this lower r and C^* to be consistent with constant average unemployment *and* with the degree of capital-deepening that capitalists wish to carry out in search of maximum transitional profits, the rate of exploitation will have to rise, in the degree indicated by the slope of the consistent path locus in the upper right quadrant. If there is a positive relationship between the unemployment rate and the rate of exploitation, the rise in the former will have the desired effect; ϵ will rise until Q^* is choked off and C^* increased until the difference $C^* - Q^* = n$. The opposite will occur if the economy is at any point to the northeast of the consistent path locus. The consistent path is, therefore, to the extent that this mechanism operates, a stable path.

Assuming the economy stays on the consistent path, it will move toward the point S at which the growth of Q comes to a halt. S represents a stable steady state in which $Q^* = 0$, and, as can be shown under certain strong conditions, the marginal productivity rule holds. In this version of Marxian accumulation, therefore, the steady state represents a final resting place toward which the system moves. In this sense the system is stable "in the large," and the question arises: How obtainable or unobtainable is the point S? The "dual" to the Marxian accumulation theorem would place the economy at S and reduce the consistent path to a historical or irrelevant exercise. It must suffice here to state that study of the coordinates of ϵ_{max} and r_{min} indicates that the point S is probably out of reach for any existing capitalist economy, and likely to remain so. It should be noted that the consistent path embodies the two laws of motion that emerge from Marx's discussion of accumulation: the falling tendency of the rate of profit and the rising tendency of the rate of exploitation.

III

In summary: capitalists in competitive conditions pursue maximum transitional, or quasi-monopoly, profit rates. In the short run, in which this pursuit alone takes place, they are constrained by production functions whose parameters are determined by an engineering practice historically adapted to capitalist production relations, and therefore emphasizing deepening over fundamental research and shifting outward. The model thus portrays an intimate connection between accumulation and technical change (in which the former tends to outrun the latter for reasons specific to capitalism), rather than accumulation *in vacuo* (as in the Cambridge post-Keynesian models) or disembodied technical change and choice without an investment function (as in the neoclassics).

The resulting relation between the growth rate of the organic composition of capital and the rate of exploitation is the capstone of the model. The derivative consistent path identifies combinations of ϵ and r that imply balance between the growth rates of the labor force and of the capital stock. The model then sets the stage for study of deviations from the consistent path and their resolution; it confirms some major Marxian predictions, such as the long-term rising rate of exploitation and slowing down of the growth rates of output and the capital stock.

Models invariably do violence to the full

texture of reality, as our best thinkers—Marx, Marshall, Keynes—all pointed out. Marx's vision included growing potential for "realization" crisis, the concentration and centralization of capital, the impoverishment of work, foreign economic-military expansion, increasingly authoritarian relations within production units, increasing reliance on the machinery of the state to reinforce reproduction at adequate levels of profit, and ultimately the cannibalization of civil society itself—and all this in a context of resistance on the part of the working class and allied strata, which pick up, one by one, the banners thrown down by the pillars of society and formulate ever-more thoroughgoing solutions. These, and more, are aspects of a Marxian theory of capitalist *development;* any model of capitalist *growth*, like good seasoning, must accent the vision from which it derives, rather than turning one's attention away from that vision.

REFERENCES

Raford Boddy and James Crotty, "Class Conflict, Keynesian Policies, and the Business Cycle," *Monthly Rev.*, Oct. 1974, *26*, 1–17.

Maurice Dobb, *Political Economy and Capitalism*, New York 1945.

F. H. Hahn and R. C. O. Matthews, "The Theory of Economic Growth: A Survey," in *Surveys in Economic Theory*, Vol. I, New York and London 1967.

Geoffrey C. Harcourt, "Some Cambridge Controversies in the Theory of Capital," *J. Econ. Lit.*, Sept. 1969, *7*, 369–405.

Nicholas Kaldor, *Essays on Value and Distribution*, London 1960.

David Laibman, "The Marxian Labor-Saving Bias: A Formalization," *Quart. Rev. Econ. Bus.*, Autumn 1976, *16*, 25–44.

———, "The Marxian Profit Cycle: A Macromodel," unpublished.

H. Lewin and J. Morris, "Marx's Concept of Fetishism," *Science & Society*, forthcoming.

Karl Marx, *Capital*, Vol. I, Chicago 1906.

James Meade, *A Neoclassical Theory of Economic Growth*, London 1961.

Joan Robinson, *The Accumulation of Capital*, London 1956.

———, *An Essay on Marxian Economics*, London 1942.

———, *Economic Heresies*, Cambridge 1972.

Robert Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Feb. 1956, *70*.

———, "Technical Change and the Aggregate Production Function," *Rev. Econ. Statist.*, Aug. 1957, *39*, 312–20.

Piero Sraffa, *Production of Commodities by Means of Commodities*, Cambridge 1960.

Trevor Swan, "Economic Growth and Capital Accumulation," *Econ. Record*, Nov. 1956, *32*.

Paul Sweezy, *The Theory of Capitalist Development*, New York 1956.

Econometric Methodology in Radical Economics

By Dale J. Poirier*

> To this day many of our comrades still do not understand that they must attend to quantitative aspect of things. . . . . They have no "figures" in their heads and as a result cannot help making mistakes. —Mao Tsetung [1949b, pp. 379–80]

Economics, whether radical or bourgeois, is concerned to a large extent with quantitative matters, and hence, it is not surprising that economic analysis turns towards statistical description and verification of abstract theorizing. This propensity towards quantitative methods is by no means a recent phenomenon. Primitive quantitative methods were evident in the tableaux of the physiocratic school and in the Malthusian population and Paretian distribution formulas. However, economists as diverse as Lawrence Klein (p. 416) and Joan Robinson (p. 76) have noted that the major impetus to the recent quantification of neoclassical economics has been Keynesian economics.

Unlike bourgeois economics, radical economics has not undergone anything akin to a "Keynesian revolution," so that comparatively speaking, it has remained largely unquantified. The main theme of this study is that there is nothing intrinsic in radical economics which precludes quantification and, hence, econometric analysis. The radical literature can be characterized in part by its paucity of empirical analysis, and while it might be argued that this has been at the expense of a wider acceptance of radical doctrine by bourgeois economists, it will be argued here that more importantly, it has been at the expense of a more sound scientific foundation for radical analysis.

*Associate Professor, Department of Political Economy, University of Toronto. Gratitude is owed to numerous individuals whose comments at various stages of development aided in preparation of this final draft; however, the opinions expressed here are the sole responsibility of the author.

I. Econometrics: Positive or Normative?

In this essay "empiricism" will be distinguished from what has been called "immanent empiricism." Martin Bronfenbrenner (p. 12) has defined immanent empiricism as the doctrine which professes that "if one looks at enough facts or cases long and hard enough, general solutions (or acceptable compromises) will become clear, less, by formal logic than by 'insight,' by 'vision,' by analogy, or sometimes by 'compulsive comparisons.' " In contrast empiricism will be used here to describe the doctrine which tempers immanent empiricism with inductive reasoning based on a body of theory. Empiricism involves the concurrent development of theory and observation, and in this sense, it is an intrinsic element of scientific inquiry. As Robert Heilbroner (p. 18) has said: "essentially the claim to being a scientific procedure rests on nothing more than a subscription to orderly repeatable methods and to the willing submission of hypothesis to empirical testing." In economics the methodology by which theory and observation are related, using appropriate methods of inference, is know as econometrics.

The important distinction between the statistician and the econometrician is that the latter employs her or his statistical tools to the analysis of *economic* models. These economic models are the products of economic paradigms, and these paradigms serve as the bases for which endogenous versus exogenous classification and identifying restrictions are made. When the researcher replaces the statistician's hat with the econometrician's hat, then these actions involve statements concerning the economic operation of the real world.

These actions serve as a source for normative inputs into econometrics. For years the debate over whether economics is a positive or a normative science has padded the publication

records of many but probably changed the minds of few. Rather than repeat all too familiar arguments here, I shall merely remind bourgeois readers of Robert A. Gordon's (p. 4) frank comment in last year's American Economic Association Presidential Address: "neoclassical economics has always had a normative slant." Radical readers, I trust, have no problem accepting a normative classification of economics. Rather than the existence of normative elements in economic theory, the issue of concern here is how these elements show up in the statistical toolbox of the econometrician.

Ultimately, the endogenous versus exogenous classification depends on the researcher's perspective and on the purposes at hand, and obviously bourgeois and radical paradigms differ in these respects. For example, bourgeois macro models treat events such as the quadrupling of oil prices by the Organization of Petroleum Exporting Countries as exogenous to the *U.S.* economy, whereas most radicals would argue that such price increases are the result of many years of *U.S.* imperialism abroad. Since in this example, as in any other, the endogenous versus exogenous classifications have direct bearing on the appropriate estimation techniques to be employed (i.e., single equation vs. simultaneous equation techniques), the choice of the appropriate estimation technique is determined within the paradigm.

Similarly, identification restrictions often reflect prejudices and value judgments. For example, suppose a model of a family's labor supply can be identified by postulating a recursive model in which the husband first determines his labor supply and then the wife determines her labor supply conditional on the husband's. Might not such a recursive formulation reflect a sexual bias which assigns a subservient role to women? Unless it is empirically supported, might not this bias lie in the mind of the researcher rather than in the real world?

More subtlely, Sidney Winter has suggested an analogy between the identification question and the "as if" theorizing principle advocated by Milton Friedman. Even Paul Samuelson (p. 232) has acknowledged that this priniciple, for which Samuelson coined the phrase "*F*-twist," is such that the "nonpositivistic Milton Friedman has a strong effective demand which a valid *F*-twist brand of positivism could supply." In the context of identification, this principle encourages the estimation of only reduced form equations since the "truth" and the "as if" theory may lead to the same reduced form. This principle is diametrically opposed to Marxian economics, which is based on an understanding of the dialectical structural form of society.

Even if endogenous vs. exogenous classifications and identifying restrictions did not provide ample opportunity for normative inputs into theoretical econometrics, applied econometrics would still provide such opportunities. Whereas the preceding two actions logically precede estimation, during estimation data massaging, despite being deplored by most econometricians, persists in applied work, and obviously the choice of models to report as the final result contains a normative element. Often the statistical significant results that arise owe more to the persistence of the researcher than to the researcher's economic insight. While under the guise of positivism many academic economics departments could be more aptly described as political oracles, the computer rooms of many institutions could be more aptly described as massage parlors.

Clarification of the normative vs. positive issue is essential to avoid confusing radicals' (justifiable) distaste for some of the normative elements reflected in the econometric tools of bourgeois economics with a distaste for the tools themselves. However, it is also important to keep in mind the simultaneous role these tools have played in the bourgeois paradigm. As Thomas Kuhn noted, one important characteristic of any paradigm is that the paradigm itself suggests the problems and questions its practitioners should investigate. While many bourgeois economists might agree that the recent increased interest in labor economics is partly responsible for the extensive research

into econometric models with limited dependent variables, they would probably balk at the claim that the line of causality ever runs in the opposite direction. Yet, for example, Klein (p. 417) has noted that the development of spectral analysis and computer simulation of dynamic systems led to renewed interest in the analysis of cyclical fluctuations. While caution might be warranted, neither this simultaneous role of statistical tools nor their use as vehicles for normative inputs in the bourgeois paradigm seem to be justifiable grounds for radicals' abandonment of these tools in their own paradigm. Thus we turn to further explanations.

II. Reasons for Radicals' Disenchantment

In 1959 Polish economist Oscar Lange observed that it was no mere coincidence that econometrics was developed in Western countries precisely during the period of growing power of monopoly and state capitalism. Indeed some of the most fundamental problems of early econometric research (e g., business cycle forecasting) were connected to the needs of economic policy of the capitalistic state. Lange, however, was also quick to point out the importance of adapting econometrics to meet the needs of socialist countries. Apparently, few Western radical economists have followed Lange's suggestion. While the reasons for this neglect are somewhat unclear, at least three possible explanations can be put forth.

First, contemporary radicals may base their stance on what has been called the "overquantification" of the social sciences—a situation over which Wassily Leontief (1966, p 46) has also expressed concern. Radicals' concern with overquantification most likely stems from a general distrust of any element of the bourgeois paradigm. As Kuhn has noted, researchers whose work is based on shared paradigms are committed to the same rules and language for scientific practice. Clearly econometric analysis has become one of the "rules" in the bourgeois paradigm, and its mathematical language resembles that of neoclassical economic theory; however, there is no reason to expect competing paradigms to be totally disjointed.

Nearly indistinguishable from this overquantification issue, is the issue concerning the emphasis of "rigor over relevance" in neoclassical economics—an issue dealt with in the Presidential Address of the *AEA* by Gordon in 1975 and Leontief in 1970. In particular, Leontief (1971, p. 3) said quite bluntly: "In no other field of empirical enquiry has so massive and sophisticated a statistical machinery been used with such indifferent results." For example, regardless of Robert Basmann's contention to the contrary, the relevance of much of the work in the past 15 years on deriving the finite sample distribution of simultaneous equation estimators, lies (at best) in the future. Econometricians cannot forever justify the relevance of their work by arguing that researchers only need wait until the rest of economics develops to the point where it will be able to use it. Such an argument is fallacious since its defenders can always argue that the rest of the profession has not waited long enough yet. However, while these overquantification and rigor-over-relevance arguments have some validity in bourgeois economics, they hardly seem justification for the general neglect of econometrics in the radical literature.

Back in 1950 Oscar Morgenstern raised the perplexing question of whether economic data will ever be sufficiently accurate to justify the need for, say, simultaneous equation estimation techniques. However, even more important than the quality of data is its existence, and as a second possible explanation for radicals' apparent adverseness toward econometrics, whether the bourgeois paradigm has had an effect on the availability of data. Clearly, the national income accounts are direct outgrowths of the Keynes revolution, and the privacy of much corporate data is, in part, the result of the "competitively vulnerable" image of the firm in neoclassical theory. More subtly, however, Duncan Foley (p. 3) has contended that in the ideological sphere pre-existing ideologies like racism and sexism are adopted, shaped, and reproduced by bourgeois society. Thus, for ex-

ample, sexism may ultimately affect the data which are available. In particular, Morley Gunderson has noted that the researcher working on labor force participation may find that data on the number of children has been collected for a sample of women, but not for men, presumably out of a preconceived notion of sexual roles.

While data shortages may confront radical researchers in some instances, in general, however, there is not a severe shortage of data—paging through a few issues of *The Review of Radical Political Economics* will uncover numerous tables of data. Unfortunately, most radical authors treat the data with extreme reverence, rather than cautious suspicion. They appear to view their data as if they were natural constants such as found in physics. Are the data that radical researchers collect so perfect that they exactly agree with an a priori, deterministic theoretical model? The answer is so obviously no, yet many radical authors compare sample estimates as if they were constants rather than realizations of random variables.

A third possible explanation for radicals' apparent disenchantment with econometrics can be described as a distrust with the second set of assumptions necessary to transform a theoretical economic model into operational form (e.g., the need to go from a concept such as "the amount of socially necessary labor embodied in a commodity" to a testable hypothesis). It seems, however, that a researcher's attitude toward econometric assumptions (e.g., normality of the error term) should be no different than her or his attitude toward the underlying assumptions of the theoretical economic model in question. The statistical assumptions of an econometric model are seldom strictly true, but rather they are usually thought to be reasonable, and hence serve as "a place to hang one's hat" and begin the analysis. The context (if any) in which to evaluate the reasonableness or realism of the assumptions depends on the purpose at hand, and hence is normatively determined.

Like any model an econometric model follows from its assumptions and may be considered in this sense to be absolutely true. The model and its assumptions, however, may not be applicable to any real world situation. The proposition that a model is reasonably applicable to a given set of situations has been called an *applicability theorem* by Bronfenbrenner. Since it relates to the real world, an applicability theorem may be highly probable, but it is never absolutely certain. Thus, for example, an econometrician's claim that a particular variable may be used as an instrumental variable is an applicability theorem. This theorem states that a particular economic variable satisfies (at least to a reasonable degree) certain statistical assumptions such as it is asymptotically uncorrelated with the disturbance term in the model and whatever economic influences the disturbance term represents.

As William Baumol has said, a researcher should use a model

> "which is sufficiently simple and orderly to be amenable to systematic study and yet which at the same time is close enough to the full facts of the matter to permit the conclusions drawn from investigation of the model to retain some relevance for the more complicated phenomenon which the model is designed to represent." [p. 90]

Econometric models, no matter how complex, are not substitutes for judgement, rather they focus attention on the factors about which judgement must be exercised. The attitude of many radical researchers toward econometric models is probably similar to Raymond Franklin and William Tabb's (p. 129) characterization of their general attitude toward economic inquiry: "They view theorizing not as an end itself nor as an aesthetic ritual involving rigorous elegance, however relevant; rather, they see it as being linked to their advocacy of fundamental change or to their analysis of the barriers which obstruct such change." While econometric models should indeed be kept in perspective to the overall goals of radical economics, they nonetheless appear to have a definite place in radical economics.

III. Bayesian and Nonparametric Analysis: Radical Tools for Radical Economists

One of the major reasons why Bayesian econometrics has not gained wider acceptance is that in most cases in order to obtain results substantively different from those obtained by classical methods, the researcher is required to express in a formal way her or his a priori subjective beliefs. While to do so may be almost sacrilegious to many believers in positive economics, radical economists should welcome the Bayesian approach with open arms since it provides a systematic way to incorporate subjective beliefs *explicitly* into the analysis. Subjective priors are based on the principle of subjective or personal probability which holds that it is meaningful to express the degree of confidence we may reasonably have in a proposition, even though we may not be able to give either a deductive proof or disproof of it.[1] Radicals who are skeptical of empiricism will have fairly "sharp" priors; however, so long as they provide some degree of uncertainty (i.e., their priors are nondegenerate), such priors permit the data to change or alter their prior beliefs, producing in the process, posterior beliefs.

Another explanation for the failure of Bayesian analysis to gain wider acceptance is that it is often impractical for large models. More often than not, however, parsimonious models are the creation of individuals who are in either the extreme lefthand or extreme righthand tails of the political spectrum since only they can narrow down their models to a small number of key variables. For example, right-wing Chicago-oriented monetarists often employ small models centered around one key variable—the growth rate of the money supply. Economists in the middle ground tend to be less certain as to what are the key variables because they are torn between numerous competing and often conflicting theories. As a result they try to take everything into account. On the other hand, while econometric models have been rather rare in the radical literature, those that have been estimated can hardly be described as having been "overfit," and hence, they would appear to be amendable to Bayesian analysis.

While it might at first appear that Bayesian analysis provides a theoretical framework by which diverse radical and bourgeois economists could reconcile their disparate a priori beliefs, such a view would be unduely naive. As Thomas Rothenberg has noted, Bayesian analysis can be useful in describing the process of normal science within a paradigm, but new paradigms are not related to old ones by a generalized Bayes formula. Bayesian analysis, or for that matter, classical statistical analysis as well, is unlikely to ever provide the impetus for proponents of a particular paradigm to undertake the dramatic abandonment of their paradigm; however, it may afford researchers, and radical researchers in particular, an inferential structure for dialectical materialism.

Besides Bayesian statistics, another area of statistics which has also not been widely accepted in the mainstream of the econometrics literature is *nonparametric statistics*. Nonparametric analysis is more robust than the conventional parametric analysis because it does not require stringent distributional assumptions for the underlying population. Furthermore, nonparametric statistics is ideally suited for the analysis of categorical and ordinal data. There is, of course, a price to be paid for this robustness in terms of the power of hypothesis tests; however, the power losses are often fairly small.

It would seem that the robustness of nonparametric approaches should be attractive to many radical researchers who do not already have vested interests in the conventional approaches of parametric statistics. Furthermore, radical researchers who are bothered by the "second level" set of assumptions that must be imposed on already simplified theoretical models in order to make them operational for statistical analysis should find the absence of stringent assumptions in nonparametric analysis quite refreshing.

[1]Note the similarity between personal probability and Michael Polanyi's concept of personal or tacit knowledge.

IV. Conclusion

This essay has examined the econometric foundations of the radical political economics literature in order to fill an apparent gap in the existing literature, and hopefully convince some radical authors to strengthen the empirical content of Marxian economics. While some may feel that this essay has been overcritical and has overemphasized the value of a sound empirical approach, the views expressed here are not all that different from those of some radical authors. In *RRPE's* recent special issue on the teaching of introductory economics, Robert Buchele and William Layonick (p. 38) left a space in their course outline and reading list and remarked "What is needed here is an introduction, from the radical economics point of view to the techniques of quantitive analysis. Such an introduction should use examples which are socially relevant, and it should outline the uses and limitations of quantitative analysis within the social scientific framework."

Unfortunately, radical researchers have not provided many such examples. Overall, both the quality and the quantity of the existing econometric work in the radical literature is rather low—although there have been occasional exceptions. The criticisms made in this essay have all been directed toward radical technique and not toward radical doctrine. Hopefully, these criticisms will all be taken in a constructive light and the words of Mao Tsetung (1949, p. 374) will be kept in mind: "We have the Marxist-Leninist weapon of criticism and self-criticism. We can get rid of the bad style and keep the good."

REFERENCES

Robert L. Basmann, "Exact Finite Sample Distributions for Some Econometric Estimators and Test Statistics: A Survey and Appraisal," in M.D. Intrilligator and D.A. Kendrick, eds., *Frontiers of Quantitative Economics, II,* Amsterdam 1974.

William J. Baumol, "Economic Models and Mathematics," in S. R. Krupps, ed., *The Structure of Economic Science,* Englewood Cliffs 1966.

Martin Bronfenbrenner, "A Middlebrow Introduction to Economic Methodology," in S. R. Krupp, ed., *The Structure of Economic Science,* Englewood Cliffs 1966.

Robert Buchele and William Layonick, "Economics as a Social Science: Introducing the Capitalist Economy," *Rev. Rad. Polit. Econ.,* Winter 1975, *6,* 20–40.

Duncan K. Foley, "Towards a Marxist Theory of Money," Institute for Mathematical Studies in the Social Sciences, Stanford University, Technical Report No. 181, Sept. 1975.

Raymond S. Franklin and William K. Tabb, "The Challenge of Radical Political Economics," *J. Econ. Issues,* 1974, *8,* 127–50.

Milton Friedman, *Essays in Positive Economics,* Chicago 1953.

Robert A. Gordon, "Rigor and Relevance in a Changing Institutional Setting," *Amer. Econ. Rev.,* March 1976, *66,* 1–14.

Morley Gunderson, "Work Patterns," In G.C.A. Cook, ed., *Opportunity for Choice: A Goal for Women in Canada,* Ottawa 1976.

Robert L. Heilbroner, "On the Possibility of a Political Economics," *J. Econ. Issues,* 1971, *5,* 1–22.

Lawrence R. Klein, "Whither Econometrics," *J. Amer. Statist Assn.,* June 1971, *66,* 415–21.

Thomas S. Kuhn, *The Structure of Scientific Revolutions,* Chicago 1970.

Oscar R. Lange, *Introduction to Econometrics,* New York 1959.

Wassily Leontief, *Essays in Economics,* New York 1966.

——— "Theoretical Assumptions and Nonobserved Facts," *Amer. Econ. Rev.,* March 1971, *61,* 1–7.

Otto Morgenstern, *On the Accuracy of Economic Observations,* Princeton 1950.

Michael Polanyi, *Personal Knowledge,* New York 1964.

Joan Robinson, *Economic Philosophy*, London 1966.

Thomas J. Rothenberg, "The Bayesian Approach and Alternatives in Econometrics II," in S. E. Fienberg and Z. Zellner eds., *Studies in Bayesian Econometrics and Statistics*, Amsterdam 1974.

Paul A. Samuelson, "Problems of Methodology Discussion," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 231–36.

Mao Tsetung, "Report to the Second Plenary Session of the Seventh Central Committee of the Communist Party of China," March 1949, in *Selected Works*, Vol. 4, first English edition, Peking.

——— "Methods of Work of Party Committees," March 1949, in *Selected Works*, Vol. 4, first English edition, Peking.

Sidney G. Winter, "Optimization and Evolution in the Theory of the Firm," in R. H. Day and T. Groves, eds., *Adaptive Economic Models*, New York 1975.

RACIAL DISCRIMINATION

A Labor Force Competition Theory of Discrimination in the Labor Market

By DAVID H. SWINTON*

The theory of labor market discrimination is concerned with how and why productively irrelevant characteristics of workers such as race influence the labor market behavior of employers and workers. The theory is an attempt to explain a wide variety of existing labor market differences between workers who differ by only productively irrelevant characteristics. Thus, the theory must offer an explanation for the residual differences in occupational distributions, wages, unemployment rates and industry or firm distributions which have been observed to exist between otherwise homogeneous workers of different races, ethnic, sex or religious groups within the same labor market. In dynamic versions, discrimination models might help explain group differences in job search patterns, investment in human captial and migration rates.

Most previous theories of racial discrimination in the labor market have been based on the optimizing behavior of employers. The employers of these models are free to choose their work force from a set of potential workers of various races. If the set of potential workers is partitioned into productively homogeneous sets in the technical sense, these partitions will not correspond to a partition by race From among these sets, the employers choose a labor force racial composition which maximizes the employer's objective function. If this racial composition differs from the composition of the partitioned labor force, then the employer is discriminating.

* State University of New York at Stony Brook

There are two classes of discrimination theories: the exogenous theories and the endogenous theories In the former case, the motivation for discrimination arises from outside the labor market. In the latter case the motivation arises from within the labor market. The labor force competition model presented here is of the endogenous variety Before we discuss the model, however, we will briefly examine the models based on employer behavior.

The exogenous models all have one common feature. Their objective function is some specification of the employer's utility function. The motivation of discrimination in these models is based on a nonpecuniary variable generally designated a "Taste for Discrimination " These models are also generally set in the neoclassical competitive environment which explains the necessity of postulating the independent racist taste The classical statement of the taste theory is Gary Becker's "Theory of Discrimination." Kenneth Arrow has further developed the analysis based on neoclassical assumptions and taste. These models however have common weaknesses; namely, they are unable to explain most labor market phenomena associated with discrimination other than wage differences and they would seem to be unstable in a neoclassical environment

Lester Thurow has used a "taste theory" in his model of discrimination but his analysis is placed in a noncompetitive environment. His model is able to explain more characteristics than the competitive model. It can also escape the instability associated with the competitive model. However, this is accomplished by total reliance on rigid assumptions concerning the nature of the distribution of skills or the flexibil-

ity of the production function with respect to the distribution of skills. This, however, is not a major fault. The major weakness of Thurow's analysis in my view is his continued reliance on the unexplained taste to drive the analysis.

The principal types of endogenous theories will only be briefly noted since other papers at this session will deal with these models. The neoclassical endogenous theory is based on the results of search and information theory (see Arrow and John McCall) In these analyses the discrimination results from the profit maximizing response of employers to uncertainty about the quality of individual workers when the real or subjective quality distributions favor the group which receives preferences.

The Marxist analysis is the other principal endogenous explanation (see e.g., Paul Baran and Paul Sweezy). These analysts also suggest profit maximization as the driving motivation. Racism is viewed as a tactic used by employers to introduce class cleavages within the working class. This tactic is intended to minimize labor cost by weakening the labor bargaining position or perhaps to stall the longer run threat to the capitalist system.

The role of employees is briefly explored by some of the neoclassical theorists. However, they in general reject employees having a significant role essentially because of their preference for competitive assumptions. In a competitive environment, as is well known, workers cannot generally affect the conditions of work. Moreover, the implications of employee "taste" for discrimination in a competitive world implies rigid segregation and no earnings differentials, which is obviously in contrast to real world labor markets. The Marxist analysis tends to downgrade the initiating role of workers because this is inconsistent with class analysis. In any case, both Marxist and non-Marxist theorists "blame" discrimination on employers.

In our model the focus of the motivation of discrimination shifts to the favored workers. In this theory, discrimination arises out of the collective attempt of workers in the dominant group to improve their economic well being. Their economic well being is improved if they are able to raise their expected incomes through discrimination. If it appears that it may be possible to improve their economic well being, then discrimination arises.

Non-Marxist economists have a long tradition of discounting the role of groups in the economic process in favor of the role of individuals. Explanations that require group action are automatically suspect. Mancur Olson's recent book on collective action clearly laid out the basis for this suspicion. In essence, the suspicion arises because individual rationality frequently conflicts with group rationality. I have argued elsewhere that several features of the discrimination case invalidate the applicability of this argument. For one thing, there are generally no positive benefits to be gained by a benefiting worker from failing to support most actions of the discriminating worker coalition. For another, exclusion may be possible from the benefits produced by the coalition. Finally, racial discrimination is a social as well as an economic activity.

In the theory of discrimination, the historical existence of the separate groups is important. Moreover, the existence of separate group identities with their implications for "we" and "they" thinking is important Racial or ethnic groups have an existence which dates prior to discrimination with the accompanying social networks of communication and patterns of social interdependency. The social capital invested in the group identities to a very important degree facilitates the exercise of discrimination.

Given the group identities, therefore, it is possible to behave collusively in exercising a strategy of discrimination. The group identities facilitate the formation of labor market coalitions over particular sets of jobs where members of the two groups come into common competition. If there is a clear-cut strategy that will lead to gains for the workers of the dominant group, they should try to impose a strategy.

The existence of a clear-cut optimum strategy

is dependent upon the existence of an imperfectly competitive labor market. In our analysis, the labor market is hierarchical. Moreover, the labor market hierarchy is stable. Thus, particular jobs are associated with particular relative incomes. Consequently, the expected incomes of workers depend upon their probability of landing different jobs. The labor market hierarchy is defined so as to lead to an internal and external hierarchy where the internal hierarchy corresponds to an occupational hierarchy and the external hierarchy corresponds to an interfirm ranking. It is also assumed that the higher-up jobs in the hierarchy contain elements of rent (i.e., the net wages are not equal across occupations). If we further assume that this hierarchy is expected to be fairly stable, workers should have a strong motivation to secure the jobs higher up in the hierarchy.

The reasons for the existence of an hierarchical labor market are complex and have been discussed extensively in the labor market literature and elsewhere by the author (see Swinton for discussion and other references). The empirical evidence certainly supports the existence of a rather stable wage structure. Some of this evidence has been summarized elsewhere by the author. Here we note just two facts: the rank correlation of earnings by occupations between 1958–69 between any two years was greater than .987 for males and greater than .9791 for females. The rank correlations between wages by industry was greater than .9121 between any two years from 1947 through 1972, and the closer the two years, the greater is the correlation. It seems to be an incontrovertible fact that there is a relatively stable structure of wages in the labor market and therefore workers should have stable expectations of gains from landing particular labor market slots.

In order for *W* workers (the dominant group) to achieve their gains, *W* workers will have to receive preference in the labor market. The *W* coalition (*CW*) therefore will be motivated to gain preference for its members. Ordinarily, the profit-maximizing employer would not give preference to *W* workers because this could result in higher costs through excess screening and perhaps through lower productivity. The empirical importance of these efficiency costs however should not be overrated as they are likely to be small given the structure of the labor market. In any case, there is no positive monetary incentive to provide preference to *W* workers if we assume homogenous workers. Thus, the workers would have to provide some incentive for the employers to give preferences. The greater the efficiency losses, the greater the required incentives.

I have not studied the process of providing incentives. Presumably the worker coalitions engage in cost raising tactics or threats of cost raising tactics to achieve their gains. Their tactics may consist of a variety of actions such as slowdowns, strikes, refusal to train *B* workers, etc. We presume that *CW* is dominant, which is taken to mean that the profitability of production in the face of a racist *CW* is greater if preference is given to *CW* members. The result will be that employers will have a pecuniary incentive to give preference to *CW* members. These preferences may under appropriate legal environments result in formal labor market agreements allocating preferred benefits to *CW*. In other instances, the agreements may be implicit and unstated but nonetheless well understood by all involved.This model describes a situation where a three-way negotiation exists between the employer and two groups of workers over the distribution of labor market benefits. The relative benefits of different labor market slots are independent of their occupancy in the hierarchy model. Thus, the two worker groups can have their positions improved by gaining a better allocation across the slots. The labor market bargaining process in this case is nonsymmetrical in the sense that in the hierarchical markets a bargain can arise between one worker group and the employers, omitting the other group. All bargaining takes place between the workers and employers. Moreover, it is also the case that the profit maximizing employer will be indifferent about the allocation of the benefits across the two worker groups so long as it does not affect profits. However, given the dominance and racist strategy of one group, the allocation of benefits to that group improves the employer's profits. Thus, the employer and the dominant group of workers can be expected to agree on a settlement which automatically fixes

the weaker group into a subordinate position.

Given, therefore, the existence of *CW* the probability of landing favored positions will be increased. The extent of the increase depends upon the power of *CW* and the inefficiencies introduced by discrimination in the obvious ways. Given these increases, it is easy to show that the expected income of an individual from group *W* increases and the expected income of an individual from group *B* decreases. Thus, ample motivation for racial discrimination can arise from the optimizing behavior of workers, given an hierarchical labor market and identifiable racial groups.

The implications of this model for the stability of discrimination revolve around two issues. The first involves the stability of the racist coalitions. The second involves the existence of economic incentives to change the outcome of the discriminatory solution.

In criticizing the logic of these group action models, the critics often ask why don't other groups such as tall people or righthanded people form up and discriminate. If either group can dominate the labor market, a similar potential for gains would exist in both of these cases. However, forming groups or coalitions requires a considerable investment in social capital to establish group identities and patterns of mutual support. As we have indicated, such group identities almost always exist prior to the development of racial discrimination. Moreover, it can be shown that given the existence of racial groups in a hierarchical labor market, no other productively irrelevant coalition basis can improve the position of *CW* group members. Thus, there is no incentive for members of the discriminating group to invest the capital in developing nonracial social groupings to serve as a basis for coalition where a clear racial group identity already exists. Clearly, if there are gains from discriminating and no gains from forming other coalitions to discriminate, the racial coalition will be a stable arrangement for conducting discrimination.

In the longer run, the desirability of discrimination on any basis might change. This would be the case if the distribution of productively relevant characteristics altered significantly in favor of the dominant group. This is likely to happen qiven the logic of human capital acquisition if a discriminatory environment continued for a long enough period or if there are strong links between previous acquisition of human captial and the current ability to attain human capital. Thus, differences in human capital may make it possible for some members of group *W* to maintain their favored slots without explicit racial discrimination In such a case, the incentive to coalesce is weakened. However, this may only be a temporary weakening if it leads to gains in the relative qualifications of the weaker group because this may once again make discrimination worthwhile. Thus, this phenomenon might well suggest cyclical fluctuations in the pattern of discrimination.

In the second case, economic incentives will exist for any party to the discriminatory solution to violate it if by so doing the party will gain economically. Clearly the *W* workers have no such incentive. Given the wage structure, the employer also has no incentive. (Thus, equal-work-for-equal-pay laws facilitate discrimination.) However, he does have an incentive to break away if by so doing he can lower labor cost by more than the cost of retaliation of *CW*. Group *B* workers are not party to the agreements but do have incentives to break them down. However, since *B* workers desire the allocations of jobs held by *CW* members because of the net wage advantages, there is a limit on their incentive to undercut prevailing wages.

In any case, whether group *B* is able to undercut prevailing wages and offer incentives to break up the *CW* employer agreement is an empirical question. Given the widespread bureaucratic procedures for wage fixing, it is likely that institutional constraints will limit such possibilities. Moreover in general we must assume that *CW* is able to command its wage and is able to protect it. If not, the discrimination will tend to break down. It seems likely, though, that given the current structure of the labor market, wage competition will be a weak source of instability.

In sum, it would appear that this model offers a fairly stable explanation of discrimination viewed as an economic process. Fundamental changes in the incentives that arise from eco-

nomic considerations require fundamental changes in the relative economic power of the two groups in the labor market, i.e., their relative abilities to affect the profitability of production. Short of such shifts, it would appear that there is little reason to expect economic incentives to lead to changes in the discriminatory solution.

However, discrimination is very much a social phenomenon as well as an economic phenomena. Although the racial group may have no obvious economic methods to change its status, there is no reason to restrict its strategy space to economic acts. In this broader sphere the group discriminated against may well undertake social and political actions to alter its subordinate economic position. Thus, a situation of general conflict may be engendered by economic discrimination and may lead to high social cost. This social conflict may actually lead to lower aggregate productive efficiency (though the economic solution is the optimum economic solution for individual *CW* employees and employers). The social cost of discrimination is thus likely to be more important in generally motivating changes in the extent of discrimination than the purely economic cost. It should also be obvious that these social costs are public in nature and thus the solution will have to be brought about by the public authorities.

This theoretical model is remarkably robust in explaining the racial characteristics of the labor market. This model is able to explain the concentration of group *B* in the labor forces of firms offering low valued employment overall and at each occupation. It explains the concentration of group *B* workers in low valued occupations relative to their qualifications. It explains the lower overall earnings of group *B* workers and the lower earnings at each firm and occupation. All of these implications have been proved by the author elsewhere. The model would also seem capable of explaining the greater unemployment of group *B* workers and the dynamic behavior of group *B* workers with respect to job search and human capital investment.

This model has not been subjected to a direct empirical test. Its implications are clearly consistent with available evidence on concentration of black workers by occupations and firms in the *U.S.* economy, the absolute and relative earnings of blacks overall and by occupation and other commonly known characteristics of the labor market. The model is also consistent with other facts such as the dissolution of discrimination in the performer side of professional athletics and entertainment. Nonetheless, since many of these facts can be explained by other models as well, it would be important to devise a specific empirical test of this model. This test would help determine how responsibility for current racial differences should be apportioned between prejudices about skills, differences in the relative stocks of human capital and current labor market discrimination of workers or psychological taste. It is likely that all of these models capture one aspect of the explanation of discrimination. This model should, however, add an additional important dimension for those interested in understanding the discrimination phenomena.

REFERENCES

Kenneth Arrow, "Some Models of Racial Discrimination in Labor Markets," in A. H Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972

Paul Baran and Paul M. Sweezy, *Monopoly Capital*, New York 1966

Gary Becker, *The Economics of Discrimination*, Chicago 1971.

John T. McCall, "Racial Discrimination in the Job Market: The Role of Information and Search," in A. H Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972.

Mancur Olson, *The Logic of Collective Action*, Cambridge, Mass. 1965.

David H. Swinton, *The Logic of Economic Discrimination in Non-Competitive Labor Markets*, unpublished Ph.D. dissertation, Harvard University 1974.

Lester C. Thurow, *Poverty and Discrimination*, Washington 1969.

Black-White Differences in Income and Wealth

By STEPHEN D. FRANKLIN AND JAMES D. SMITH*

This paper presents results from an unusual microdata set assembled by the authors and researchers at the Social Security Administration. The data set pools information from three sources: death certificates for residents of Washington, D.C. dying in 1967; Washington, D.C. estate tax returns; and Social Security earnings records. Under an arrangement worked out by Smith with the city of Washington and the National Center for Health Statistics, all (about 2,500) estate tax returns for 1967 decedents were matched with their death certificates The match provided information on age, sex, race, place of birth, marital status, cause of death and assets and liabilities. Washington, D.C. has its own estate tax, which unlike the federal estate tax, starts at a very low ($1,000) filing level A full description of this part of the data base and an estate multiplier estimate of the distribution of wealth in Washington, D C has been published elsewhere (Smith). This year, thanks to our colleagues, Frederick Scheuren and Wendy Alvey of the Social Security Administration, a procedure was worked out which permitted us to turn over to them our files and to obtain from them analytical results from matched records from our files and their records of covered earnings under the Social Security Act.

The intended use of this data base is to estimate a lifetime savings model with earnings as a key determinant. We still may be able to do so, but the prospects look rather grim. In the spirit that science is advanced by knowing what doesn't work as well as what does, we present below a few initial findings which show some promise and of a lot of statistical husbandry which bore little fruit. We shall proceed by first looking at differences in the levels of covered income reported by black and white workers, then at the wealth levels of blacks and whites, and finally at an attempt to predict the wealth of black and white workers using demographic variables and earnings records.

I. Black-White Differences in Earnings

In Figure 1 the results of an Automatic Interaction Detection *(AID)* analysis of covered earnings is portrayed. *AID* is a statistical technique developed by John Sonquist and James N. Morgan and later enhanced by Sonquist, Elizabeth Baker and Morgan. Basically, the algorithm takes a population and splits it into two groups on the values of a set of predictor variables supplied to it by the user. The combined variance around the mean (or a regression line) of each resultant subpopulation pair is computed and compared to the variance of each other possible subpopulation pair to determine which values of which variable split the initial population into the subpopulations that have the smallest combined variance. This process continues, splitting each subpopulation into two smaller subpopulations until additional splitting will not reduce the original variance by some minimum proportion (.3 percent was used here) specified by the user or until any split would result in a group with a number of observations less than the minimum (25 was used here) specified by the user.

Because the Social Security records with which we worked stretch back to 1951 and the maximum level of covered earnings has changed several times over the period, a relative measure of income was constructed, which is a worker's recorded earnings (up to the maximum) in a year divided by the maximum level of earnings taxed under Social Security in that year. Thus a

*The Urban Institute and Yale University and the Urban Institute, respectively

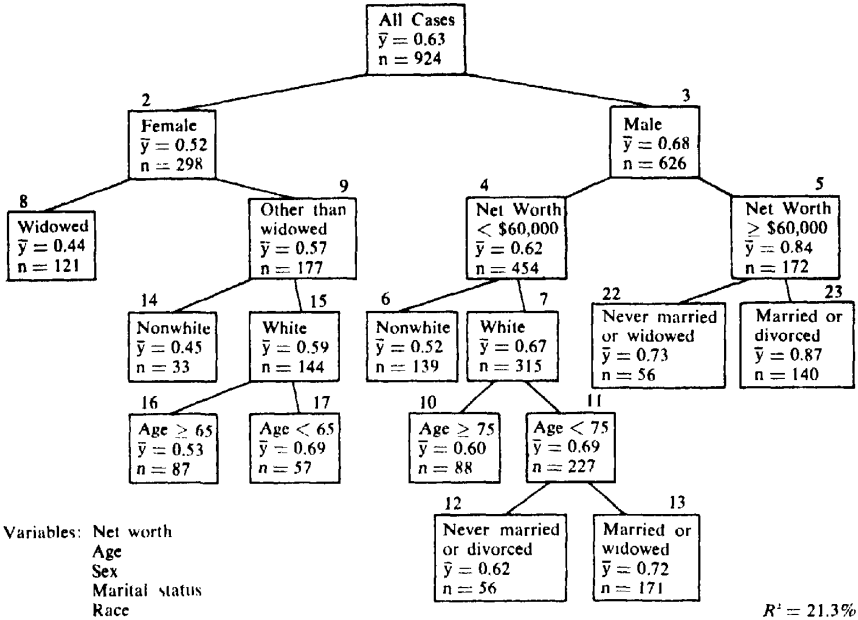


FIGURE 1 PREDICTORS OF AVERAGE EARNINGS/MAXIMUM TAXABLE EARNINGS

worker with the maximum taxable earnings in a year would have 100 percent of the maximum in that year. In order to take advantage of the repeated observations on workers' earnings, an average annual earnings divided by maximum taxable earnings was computed for each worker with four or more nonzero earnings years.

In Figure 1 it can be seen that the 924 individuals with four or more years of nonzero covered employment had an average percent of maximum covered earnings of 63 percent. The algorithm found that by splitting the population into male and female subpopulations, the greatest reduction of variance around the mean could be obtained from the variables available to it. Male workers (group 3) had an average percent of maximum covered earnings of 68 percent while females (group 2) had an average of 52 percent. Among females it was found that marital status further reduced variance in the percent of covered earnings maximum with widows averaging 44 percent and other marital groups averaging 57 percent. Among women who were other than widowed, race made a difference in expected earnings: white women (group 15) averaged 59 percent of the maximum while nonwhite women (group 14) averaged only 33 percent of the maximum. Although age could further explain the variance in the percent of maximum earnings received by white women, none of the available variables could reduce the variance in the dependent variable for black women who were not widowed (group 14).

Among men, net worth was positively associated with the percent of maximum covered earnings received. Those with $60,000 or more averaged 84 percent of maximum earnings and

those with less than $60,000 averaged 62 percent of the maximum. For the richer men ($60,000 or more) marital status was the only available variable which could further reduce variance around the mean percent of maximum, with married and divorced men showing an average percent of maximum covered earnings of 87 percent.

For men with net worth less than $60,000 at death, whites had received an average of 67 percent of the maximum and nonwhites only 52 percent. Some further reduction in variance is explained among white men by age and marital status.

The central fact to emerge from Figure 1 is that once race emerges as an explanatory variable, no further reduction in variance can be obtained by splitting the nonwhites within the constraint that a split must explain at least .3 percent of the initial variance around the mean of the dependent variable, and that a split cannot result in a group with fewer than 25 observations. Nonwhite women who were not widows had the lowest average percent of maximum earnings, 33 percent; and nonwhite men with net worth less than $60,000 had the next lowest percent of maximum earnings, 52 percent

II. Net Worth

The net worth of whites in Washington, D.C. in 1967 was on average about 19 times that of blacks, $19,300 compared to $1,000. Not only was the wealth position of blacks considerably below that of whites, but also the composition of black wealth was considerably different from that in the portfolios of whites. In Table 1 the distribution of assets and the composition among whites and blacks is shown. It can be seen there that black wealth is much more concentrated in residential real estate while only about one-fifth of white wealth is so held. White wealth-holders, and white males, in particular, tended to hold a large share of their wealth in corporate stock while blacks held practically none of their wealth in this form. In part, differences in composition reflect differences in total wealth, but even after adjustments are made for wealth level, substantial taste differences remain.

III. Estimation of Net Worth

As noted in the introductory comments, it was anticipated that earnings histories would be powerful predictive variables for estimating net worth. Using 924 cases, for whom we had at least four year's observations of nonzero covered earnings, *AID* was used with net worth as a dependent variable. The results of two *AID* runs are shown in Figures 2 and 3. In the first *AID* run average annual earnings was used as a predictor along with age, sex, marital status and race. Altogether, these variables were able to explain only 2.5 percent of the variance of net worth. Average covered earnings turned out to be the most effective variable, explaining .9 percent. Persons with average earnings less than $4,000 had a net worth of $47,250, those with earnings of $4,000 or more had an average net worth of $236,316. Age and marital status also were powerful enough to enter into the prediction, and together added on 1.6 percentage points to the reduction in variance

Because we were dealing with income measures over a long span of years, it was feared that the meaning of an average income computed in the early years of the period would not be the same as an average computed for the later years. Consequently, the average percent of maximum covered earnings concept described earlier was substituted into the *AID* analysis in place of the average earnings concept. When this was done a slightly improved R^2 was obtained: 2.8 percent compared to 2.5 percent. However, the new predictor, average percentage of maximum covered earnings, explains 1.8 percent of the variance in net worth, or twice as much as average earnings. The only other variable capable of reducing variance was age, with those over age 65 having the larger net worth.

It should be noted that race does not show up as a powerful predictor in either of the last two *AID* analyses. This occurs because low earnings and blackness are correlated and low income is a slightly better predictor. Once a split occurs on average income, all the black sample members end up in the low net worth group. In Figure 3, for instance, group 2 contains 200

TABLE 1—COMPOSITION OF WEALTH BY SEX FOR BLACKS AND NONBLACKS WITH $1,000 OR MORE NET WORTH IN WASHINGTON, D.C., 1967
(percentage of net worth)

	All Racial Groups			Blacks			Nonblacks		
Asset	Total	Male	Female	Total	Male	Female	Total	Male	Female
Real estate	22.7	20.3	25.3	81.5	84.9	76.5	18.2	14.7	22.0
Stocks and bonds	32.6	56.8	5.1	5.2	3.6	7.6	34.7	61.5	5.0
Notes, mortgages, cash and deposits	20.7	19.5	22.1	23.8	22.3	26.2	20.4	19.2	21.8
Miscellaneous	8.3	10.7	5.6	11.7	15.1	6.4	8.0	10.3	5.5
Pension funds	2.5	3.4	1.5	3.9	4.3	3.3	2.4	3.3	1.4
Power of appointment	0.3	[a]	0.6	[a]	[a]	[a]	0.3	[a]	0.6
Lifetime transfers	3.5	4.3	2.5	2.8	4.0	1.0	3.5	4.4	2.6
Gross assets	110.7	115.1	105.8	128.9	134.3	120.9	109.4	113.4	104.8
Debts	10.7	15.1	5.8	28.9	34.3	20.9	9.4	13.4	4.8
Net worth	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Life insurance[b]	11.9	18.1	4.3	26.1	39.3	6.4	12.0	17.0	4.1
Joint property[b]	14.1	6.5	22.9	52.9	59.8	42.7	19.9	18.3	21.6

[a] Rounds to less than .1 percent
[b] Life insurance and joint property are not included in the concept of net worth, but are shown here as information items.

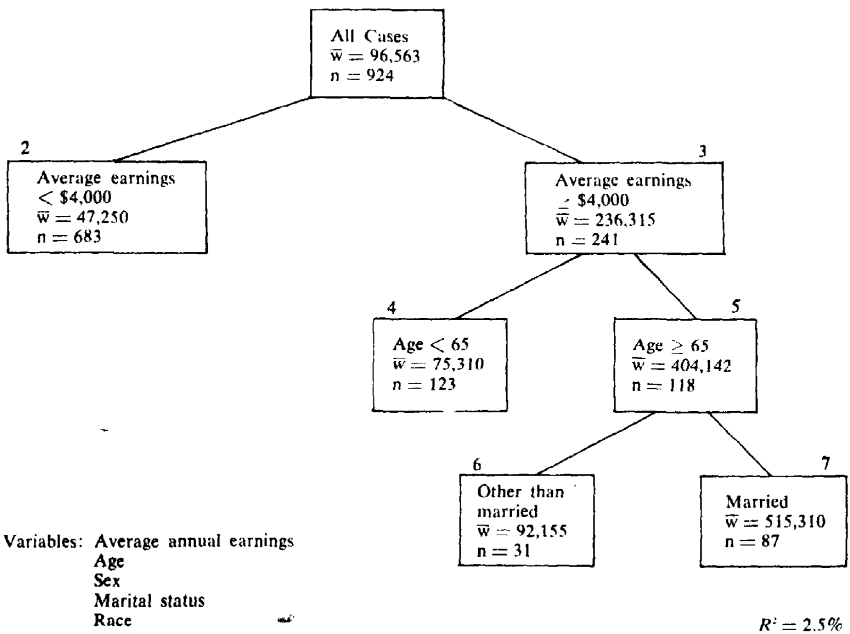
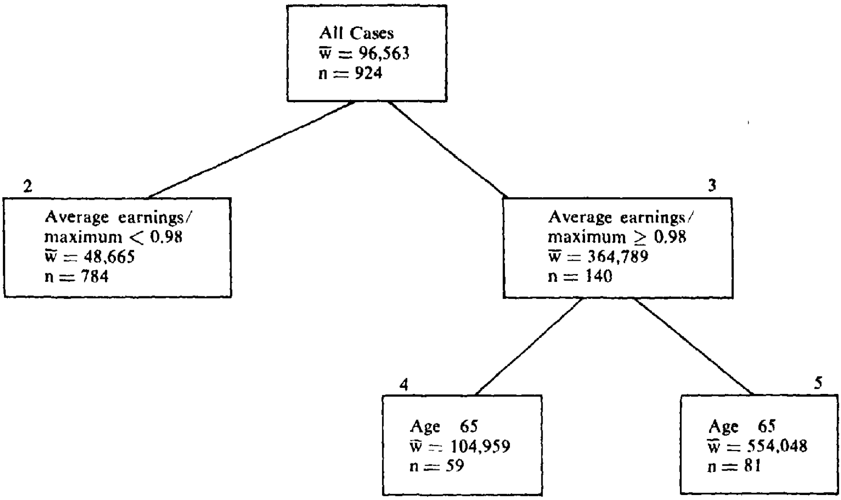


FIGURE 2. PREDICTORS OF NET WORTH



Variables: Average percentage of maximum taxable earnings
Age
Sex
Marital status
Race

$R^2 = 2.8\%$

FIGURE 3. PREDICTORS OF NET WORTH

black individuals and group 3 only 9.

IV. Conclusions

Although attempts to explain net worth in microdata have generally been of limited success, the availability of microdata earnings histories suggested a possibility for a substantial improvement. Measured against past efforts the results so far are disappointing; measured against our anticipations, they are miserable. The cut-off of earnings at the Social Security maximum is a serious data deficiency. Also, the large number of decedents in the Washington, D.C. area who are not in covered employment under the Social Security Act limit the utility of the sample.

Other avenues which suggest themselves are matching estate tax returns with income tax returns. Tax returns have the advantage of capturing higher incomes and including in the sample individuals whose employment is not covered under the Social Security Act.

REFERENCES

James D. Smith, "White Wealth and Black People: The Distribution of Wealth in Washington, D.C., 1967," in *The Personal Distribution of Income and Wealth,* National Bureau of Economic Research 1975.

John Sonquist, Elizabeth Baker and James N. Morgan, *Searching for Structure,* Institute for Social Research, University of Michigan 1973.

SELECTED CONTRIBUTED PAPERS

Wives' Labor Force Behavior and Family Consumption Patterns

By MYRA H. STROBER*

In 1940 the labor force participation rate for married women, husband present, was 14 percent. By 1970 it had increased 26 percentage points to 40 percent. The supply and demand variables associated with this increase have been widely investigated. However, there has been very little research on the economic *effects* of wives' labor force participation. (See R. Agarwala and J. Drinkwater, Margaret Carroll, Robert Holbrook and Frank Stafford, Lucy Mallan and Jacob Mincer.) This paper analyzes two economic effects: the effect on the ratio of consumption to income (C/Y) and the effect on the ratio of durable goods purchases to income (Dur/Y). The primary question addressed is: Controlling for total family income and several other variables, what are the differences (if any) in the ratios of C/Y and Dur/Y for working-wife $(W\text{-}W)$ and nonworking wife $(N\text{-}W\text{-}W)$ families? Looked at in another way, the question may be rephrased: How do *W-W* families use wives' income?

I. A Theoretical Framework and Two Hypotheses

Although based on quite different assumptions and paradigms, the two theories extant in the literature which deal with the relationship between wives' labor force behavior and family consumption patterns both postulate that, total family income held constant, the C/Y ratio will be lower in *W-W* than in *N-W-W* families. John Kenneth Galbraith's conclusion flows from combining the observation that family consumption requires highly labor-intensive consumption administration with the suggestion that working wives have less time for (and possibly less interest in) such administration. Jacob Mincer's deduction, on the other hand, is reached by incorporating into Milton Friedman's permanent income theory the assumption that wives' earnings have a large transitory component.

While Galbraith and Mincer reach the same conclusion with respect to the C/Y ratio, they propose opposite hypotheses with respect to the Dur/Y ratio. Galbraith posits a lower Dur/Y ratio for *W-W*, as compared with *N-W-W* families. However, since Mincer regards the purchase of durables as a form of saving, he hypothesizes that, holding total family income constant, the Dur/Y ratio will be higher for *W-W* families.

Neither of these theories is sufficient to fully explain relationships between wives' labor force behavior and family consumption patterns. Galbraith's theory ignores the important fact that some durables can save time and effort even after consumption administration requirements are accounted for. Mincer's thesis, because it relies on the assumption that families treat wives' earnings as transitory income, also seems somewhat narrow. It appears to me that

*Assistant Professor of Economics, Graduate School of Business, Stanford University The research on which this paper is based was supported in part by the Stanford University Research and Development Fund and in part by a grant from the E. I. du Pont de Nemours Co to the Stanford Graduate School of Business. I wish to thank Alice Amsden, Michael Cummins, William Dunkelberg, Robert Flanagan, Martin Rein, Robert Michael, Frank Stafford, Charlotte Stiglitz and Robert Willis for helpful discussions and Lynn Rosener for research assistance

the following theoretical framework, which builds upon the work of James Duesenberry, provides a more useful approach to explaining the relationship between wives' labor force behavior and family consumption.

For most wives, the economic motivation to work is closely associated with husbands' earnings. Most families have a life-cycle reference group with whom they compare themselves. This reference group tends to be similar in age, education level, geographic region, etc. When a *N-W-W* family finds a gap between its income and consumption levels and those of its reference group, the wife in that family is likely to work some number of hours so that, at the relevant wage rate, sufficient income to close the income-consumption gap is obtained.

In families where wives have very low market wage rates, wives' labor supply may not be forthcoming even though there exists an income-consumption gap with respect to life-cycle reference groups. In such situations, the effort required by the wife to eliminate or significantly reduce the gap is deemed too arduous and/or not worthwhile in terms of lost home production. On the other side of the income distribution, among wives with high levels of education and hence relatively high earning ability and "tastes" for work, the level of husbands' earnings or the concept of an income-consumption gap may play only a minor role in determining labor supply. However, for the vast majority of families, wives work in order to raise their family incomes to those of their life-cycle reference group.

In an effort to become like their friends and neighbors, *W-W* families by and large plan to use wives' incomes to purchase durables, nondurables and services and to save in approximately the same proportions as *N-W-W* families with the same total income. However, once the wife is in the workforce, *W-W* families find that although they may have the same aggregate income as their *N-W-W* family counterparts, they are in fact quite different in several respects from these other families. These differences cause *W-W* families to have a higher C/Y ratio than *N-W-W* families, but to maintain parity with respect to the Dur/Y ratio.

What are the key differences between *W-W* and *N-W-W* families? First, and foremost, because *W-Ws* perform considerable amounts of housework in addition to their market work, their total work week, market plus nonmarket work, is significantly longer than that of *N-W-Ws* (about 11 hours longer for women employed full-time. See Joann Vanek.) Thus, for a given vector of prices facing consumers, *W-Ws* should find it more profitable to substitute time-saving (and probably also fatigue-saving) goods and services for home production. However, most families already own such time-saving durables as refrigerators and stoves and many also own washers, dryers and dishwashers. Moreover, time and effort-saving durables purchases, while expensive, tend to be nonrecurring. Thus, although initial labor force participation by wives may be associated with an increase in the Dur/Y ratio, after wives have been at work for a few years, most of the substitution out of home-production is likely to be into time-saving nondurables (e.g., convenience foods) and services (laundries, restaurants, etc.).

The second major difference between *W-W* and *N-W-W* families is that *W-Ws* incur work-related expenses (transportation, clothing, child-care, etc.). Thus, total work related expenditure, and, therefore, total consumption, is likely to be greater in *W-W* than in *N-W-W* families with comparable incomes.

The final difference between the two types of families concerns the motivation to save. Having a working wife may well diminish a family's motive to save as a hedge against husband's job loss. Each earner tends to lessen the need for reliance on savings should the other become unemployed or disabled. Moreover, if a working wife is covered by a pension plan, which is in part employer financed, the family's motive to save for retirement may also be lessened. The strength of the relationship between wife's employment and reduction in family motivation

to save is likely to be a positive function of the proportion of total family income earned by the wife.

In summary, the theory proposed here leads to the following two hypotheses: total family income held constant, (1) the C/Y ratio will be higher in *W-W* families than in *N-W-W* families but (2) the Dur/Y ratio will not differ across the two types of families.

II. Data and Models

The data for this study are panel data from the Michigan Survey Research Center 1967–70 Survey of Consumer Finances. I concentrate here on the data for 1968 for families with husbands between the ages of 25 and 64. In order to test more accurately the two hypotheses, the sample excludes non-husband-wife households; farmers and farm managers; families where the husband was retired, permanently disabled, or a student and families who received inheritances in 1968.

Total consumption (C) was not directly measured in the survey; it is derived, as noted in equation (1), by adding to 1968 total family income (Y) the change in debt from 1967 to 1968 and then subtracting the change in savings (S) from 1967 to 1968 as well as 80 percent of expenditures on mortgages *(Mortg)*, net outlay on durables *(Dur)* and net outlay on automobiles *(Cars)*. Durables are defined as furniture, refrigerators, washing machines, stoves, television sets, household appliances and air conditioners.[1]

$$(1)\qquad C_{68} = Y_{68} + (Debt_{68} - Debt_{67}) - (S_{68} - S_{67}) - .8(Mortg_{68} + Dur_{68} + Cars_{68})$$

Consumption is hypothesized to be a function of current family income (Y), human and nonhuman wealth, life-cycle stage, expectations about future income, and wife's labor force behavior. Net assets *(NASTS)*, assets minus debt, measures nonhuman wealth, where assets equal total savings in 1968 plus the undepreciated value of cars and durables in 1968. A variable created by adding husband's and wife's levels of education *(HWEDUC)* measures both taste for current vs. future consumption and human wealth. For each spouse the variable is scaled as follows: 1 = 0–5 grades; 2 = 6–8 grades; 3 = 9–11 grades; 4 = 12 grades; 5 = 12 grades plus other noncollege training; 6 = some college, no degree; 7 = college, Bachelor's degree; 8 = college, advanced or professional degree. Thus, *HWEDUC* ranges from 2–16. Life-cycle stage is measured by three dummy variables (*H*35–44, *H*45–54, and *H*55–64), each taking on a value of 1 if the husband is respectively 35–44, 45–54, or 55–64. Expectations about future income is measured by a dummy *(EXHIEARNS)* which takes on a value of 1 if the husband is less than 45 and in a professional or managerial occupation. Thus, the basic consumption regression equation is,

$$(2)\quad C = a + b_1 Y + b_2 NASTS + b_3 HWEDUC + b_4 H35\text{–}44 + b_5 H45\text{–}54 + b_6 H55\text{–}64 + b_7 EXHIEARNS$$

where b_1, b_2, and b_7 are expected to be positive b_3, b_4, b_5 and b_6 to be negative.

The significance of wife's labor force behavior on consumption is measured in four ways. First, I add to (2) a dummy (*WDIN*68) which is equal to 1 if the wife worked at least one hour during 1968. If this dummy is positive and significant and does not change the coefficient on Y, then I conclude that, *cet. par.*, *W-W* families have a higher C/Y ratio than *N-W-W* families. A second test is performed by substituting the continuous variable, hours worked in 1968 *(HRSYR)*, for *WDIN*68 and applying the same test to that variable. It is important that b_1 not change when *WDIN*68 or *HRSYR* are entered in order to be somewhat

[1]Total family income (Y) excludes both realized and nonrealized capital gains. Debt is defined as the sum of mortgage debt, installment debt, amount owed on stocks and real estate and miscellaneous debt. Savings is defined as the sum of amounts in checking and savings accounts, certificates of deposit, and the value of stocks, bonds and real estate. Expenditures on mortgages *(Mortg)* is the annual mortgage payment. Net outlay on durables *(Dur)* and net outlay on automobiles *(Cars)* are exclusive of finance charges and are defined as the price of durables (or autos) minus the value of any trade-in.

assured of the lack of collinearity between either of these variables and Y. (The correlation between Y and $WDIN68$ is .006; between Y and $HRSYR$, .08.)

The third and fourth methods of testing the significance of wife's labor force behavior on consumption are quite similar. The third method consists of substituting two separate income variables for Y in equation (2), one equal to $Y \times WWF$ and one equal to $Y \times NWWF$, where WWF is equal to 1 if the wife worked in 1968 and $NWWF$ is equal to 1 if the wife did not work in 1968. The regression using these two income variables is then compared with (2) by means of a Chow test and the resultant F is tested for significance. Test four involves substituting three variables for Y in equation (2): wife's earnings (Y_W); total family income minus wife's earnings, called other family income (Y_{OF}); and an interaction term ($Y_W \times Y_{OF}$) The regression using these three income variables is then again compared with (2) by means of a Chow test and examined for significance.

The regression with net outlay on durables as the dependent variable contains the seven independent variables in the consumption regression plus one additional variable. On the assumption that a recent change of residence is highly positively related to the purchase of durables, I create a dummy, $MOVHSREC$ which is equal to 1 if the family moved into a different apartment or home in 1967 or 1968. Precisely the same four procedures as those described above are applied to test the significance of wife's labor force behavior on net outlay on durables.

An alternative two-stage specification of the consumption and durables regression was also attempted. However, in the stage-one regression estimating wife's hours worked, the R^2 was so low (.06) that the attempt was abandoned.

III. Results

A. *t-Tests*

Table 1 presents data for 433 W-W and 379 N-W-W families with husbands 25–64 and for four life-cycle subgroups, based on husband's age. Student t-tests are employed to test the significance of differences in means between W-W and N-W-W families. The most striking aspect of the table is that, while in the aggregate and for each life-cycle group, the means of other family income (Y_{OF}) are significantly higher for N-W-W families, the means of total family income (Y) and also of disposable family income (Y_D) are the same for the two family groups. (Y_D was not directly measured by the Michigan Survey, but was estimated for each family by the Center staff.) As hypothesized, wives' earnings tend, on the average, to equalize the incomes of W-W and N-W-W families.[2]

The second interesting finding is that the means of the variables C/Y and C'/Y_D (where, in calculating C', Y_D replaces Y in (1)) are higher in W-W than in N-W-W families in every life-cycle group. In the aggregate and for the group 35–44 this difference in means is significant at the 5 percent level. The means of the variables Dur/Y_D and Dur/Y_D where $Dur > 0$ are, on the other hand, not significantly different either in the aggregate or in any of the life-cycle groups.

Several other differences and non-differences between the two sets of families are also noteworthy. N-W-W families have significantly greater mean net assets than W-W families, while W-W families have more cars and spend more on cars On the other hand, between W-W and N-W-W families, there are virtually no differences in mean debt, the mean ratios of vacation expenditures/Y_D, hobby expenditures/Y_D, or college education expenditures/Y_D.

B. *Consumption Regressions*

The results for the consumption regressions

[2]If reference groups are defined by husbands' education level rather than husbands' age, the hypothesis that wives' earnings equalize family income within a reference group is substantiated only for middle and higher education groups (where husbands education is $\geq$ twelve years) In the reference group where husbands have less than twelve years of schooling, there is a significant difference in mean total family income but no significant difference in other family income. Thus, for this education group wives' earnings improve rather than equalize family income.

TABLE 1—VARIABLE MEANS FOR *W-W* AND *N-W-W* FAMILIES

	I *H25–34*		II *H35–44*		III *H45–54*		IV *H55–64*		Total *H25–64*	
	W-W ($N = 122$)	*N-W-W* ($N = 103$)	*W-W* ($N = 136$)	*N-W-W* ($N = 102$)	*W-W* ($N = 125$)	*N-W-W* ($N = 117$)	*W-W* ($N = 50$)	*N-W-W* ($N = 57$)	*W-W* ($N = 433$)	*N-W-W* ($N = 379$)
Other Family Income (Y_{OF})	$8,236	10,328*	10,611	13,064*	10,333	14,344*	8,556	11,802*	9,624	12,527*
(Total Income Less Wives' Earnings)	(330)[a]	(540)	(453)	(882)	(692)	(962)	(624)	(1090)	(276)	(445)
Total Family Income (Y)	$10,581	10,328	13,603	13,064	13,873	14,344	11,927	11,802	12,636	12,527
	(365)	(540)	(510)	(882)	(879)	(962)	(744)	(1090)	(335)	(445)
Disposable Income (Y_D)	$9,401	9,191	11,854	11,339	11,888	12,220	10,307	10,106	10,994	10,842
	(296)	(427)	(402)	(623)	(593)	(709)	(602)	(830)	(244)	(329)
Consumption/Income (C/Y)[b]	72	56	81	.45**	73	.64	89	77	77	.59**
	(.047)	(145)	(074)	(148)	(068)	(.115)	(.083)	(279)	(035)	(.078)
Consumption/Disposable Income (C'/Y_D)[c]	69	51	79	41**	.69	.60	.87	73	.74	.54**
	(053)	(.158)	(085)	(154)	(079)	(138)	(095)	(.322)	(039)	(088)
Durables/Disposable Income (Dur/Y_D)	.0382	.0332	0309	0240	0202	0226	.0205	0124	0287	0243
	(005)	(.003)	(004)	(003)	(.003)	(003)	(005)	(003)	(.002)	(002)
Durables/Disposable Income (Dur/Y_D); $Dur > 0$[d]	0598	.0496	0489	.0437	.0382	.0433	0411	0371	.0487	.0449
	(006)	(.004)	(005)	(004)	(004)	(004)	(007)	(005)	(003)	(.002)
Net Assets	$8,196	10,026	18,878	23,567	19,214	37,595*	25,501	41,796	16,730	26,959*
	(627)	(1222)	(2392)	(2945)	(1821)	(7224)	(3402)	(9133)	(1046)	(2816)
Debt	$9,321	9,565	13,128	11,370	7,882	10,088	5,816	2,570	9,697	9,160
	(810)	(898)	(1749)	(1675)	(922)	(2074)	(1410)	(687)	(681)	(836)
Number of Cars	1.42	1 22*	1 55	1 46	1 63	1.52	1.52	1 35	1 533	1 398*
	(.053)	(0.58)	(057)	(065)	(069)	(073)	(071)	(.099)	(032)	(036)
Net Outlay on Cars/Disposable Income (Car/Y)	0741	0635	0621	.0705	0770	.0617	0536	.0838	0688	0679
	(011)	(011)	(009)	(011)	(010)	(010)	(.013)	(018)	(005)	(006)
Vacation/Disposable Income (V/Y_D)	0118	0122	0149	0183	0137	018	0152	.0154	0137	.0161
	(002)	(002)	(002)	(004)	(002)	(003)	(003)	(003)	(001)	(.002)
Hobby & Recreation Items/Disposable Income ($H\&R/Y_D$)	0048	0083	0135	.0096	0089	0064	.0040	.0053	.0086	0076
	(.001)	(002)	(006)	(003)	(003)	(002)	(002)	(003)	(.002)	(.001)
Expenditures on College Education/Disposable Income ($Educ/Y_D$): $Educ > 0$	—	—	—	—	—	—	—	—	1372[e]	.1299[f]
	—	—	—	—	—	—	—	—	(013)	(.013)

*Indicates difference in means is significant at the 1% level.
**Indicates difference in means is significant at the 5% level.
[a]Numbers in parentheses are standard errors of the means
[b]Mean consumption/Income is not the average propensity to consume (APC) $APC = \bar{C}/\bar{Y}$ Mean consumption/Income $= \frac{\sum_{i=1}^{n} C/Y}{N}$
[c]Consumption here is $= C' = Y_D + \Delta Debt - [\Delta S + 8\ (Mortg + Dur + Cars)]$
[d]The sample sizes for this variable are as follows I W-W = 78; I N-W-W = 69, II W-W = 86, II N-W-W = 56, III W-W = 66, III N-W-W = 61; IV W-W = 25, IV N-W-W = 19, Total W-W = 255, Total N-W-W = 205
[e]$N = 54$.
[f]$N = 33$

are presented in the top half of Table 2. In regression *A*, b_1, the marginal propensity to consume, all else being constant, is low, .437, and is significant at the 1 percent level. The marginal propensity to consume *(MPC)* out of net assets all else constant is .148 and is also significant at the 1 percent level. The only other significant variable is *HWEDUC*. The mean education level for the sample is 8.5, about 12 years of education for each spouse. All else constant, an additional "unit" of education for either spouse (units having been defined in Section II) decreases consumption by about $725. (Mean consumption was $8,787.) This may be the result of more patience on the part of more educated persons or of a difference in tastes (e.g., college education for children) which requires higher savings rates. The adjusted R^2 ($\bar{R}^2$) for regression *A* is .148.

I also examined possible effects on consumption of number and age of children,[3] race, husband's unemployment, and having a large decrease or increase in income in 1968 as compared with 1967. None of these variables was significantly related to consumption or to net outlay on durables and none increased $\bar{R}^2$. Nor did adding 23 families with heads under the age of 25 significantly affect the regression coefficients. Changing the definition of consumption to include as saving 20 percent of mortgage payments, net outlay on durables and net outlay on cars (rather than 80 percent) raised the *MPC* to .468 but otherwise did not affect any of the regression coefficients.

In regression *B* (Table 2) the dummy variable, *WDIN*68, which is equal to 1 if the wife was employed in 1968, is significant at the 1 percent level and indicates that, all else constant, having a working wife in 1968 raised total consumption by $3,600. The coefficient on *Y* changes slightly, to .409, when the dummy is introduced. In regression *C*, the variable *HRSYR* is significant at the 1 percent level and indicates that, all else constant, an additional hour of work above the mean (659 hours) increases consumption by $2.08. The coefficient on *Y* is virtually the same as in *A*. It is not clear whether the small change in b_1 in regression *B* meets the test that coefficients not change when *WDIN*68 is introduced. I am inclined to accept the significance of both the dummy and the *HRSYR* variables but to be cautious about placing credence in their size.

When regressions *D* and *E* are compared with regression *A* by means of Chow tests, the hypothesis that either *D* or *E* is the same as *A* may be rejected at the 1 percent level (for *D*, $F_{2,798} = 9.3$; for *E*, $F_{3,797} = 5.2$.)

C. *Durables Regressions*

In the durable regression *A′* (see bottom part of Table 2), $\bar{R}^2 = .159$. The sample sizes for the durables regressions are slightly larger than those for the consumption regressions because there were about 80 families with missing observations on the consumption variable. The most significant variable in regression *A′* is the dummy, *MOVHSREC*, which indicates that, all else constant, the 7 percent of all families who changed their place of residence in 1967 or 1968 spent an additional $437 on durables in 1968. The mean outlay on durables for the sample was $280. Size of income is also a significant determinant of durables expenditure; the marginal propensity to consume durables, all else constant, is .013. *NASTS*, *HWEDUC*, *H*34–44 and *EXHIEARNS* are not significant. However, *H*45–54 has a significant negative coefficient of $99 and *H*55–64 a significant negative coefficient of $123. Thus, while life-cycle group does not appear to be negatively related to consumption expenditures *in toto*, it is significantly negatively related to durables expenditures.[4]

[3] In regressions substituting age-of-children dummy variables for life-cycle dummy variables, the results were quite similar to those reported in Table 2. Age-of-children dummies were insignificant.

[4] In regressions substituting age-of-children dummy variables for life-cycle dummy variables, none of the age-of-children dummies were significant. Nor was number of children significantly related to durables expenditures.

TABLE 2—COEFFICIENTS FOR INDEPENDENT VARIABLES IN CONSUMPTION AND DURABLES REGRESSIONS[a]

Regression	Sample Size	$\bar{R}^2$	a (constant)	b_1 (Y)	b_2 (*NASTS*)	b_3 (*HWEDUC*)	b_4 (*H35–44*)	b_5 (*H45–54*)	b_6 (*H55–64*)	b_7 (*EXHIEARNS*)	*WDIN68*	*HRSYR*	$Y \times WWF$	$Y \times NWWF$	Y_w	Y_{OF}	$Y_w \times Y_{OF}$	b_8 (*MOVHSREC*)
								Consumption[b] Regressions										
A	808	.148	6515 94	437*	148*	−725 71*	−1034 99	−1206 84	−2565 18	3270.29	—	—	—	—	—	—	—	
				(109)	(019)	(263 27)	(1765 01)	(1854 64)	(2282 58)	(2033 79)								
B	802	154	4811 76	409*	156*	−744 86*	1176 08	−1062 94	−2274 24	3784 53	3597 77*	—	—	—	—	—	—	
				(110)	(020)	(264 81)	(1768.88)	(1858 20)	(2296 51)	(2052 38)	(1333 67)							
C	795	179	6143 18	431*	162*	−857 61*	−1414 56	−813 51	−2996.34	4038 47*	—	2 08*	—	—	—	—	—	
				(107)	(019)	(255 20)	(1715 37)	(1797 29)	(2210 06)	(1976 99)		(778)						
D	802	.162	6668 68	—	169*	−763 14*	−1425 84	−1233 47	−2555 53	3889 18	—	—	582*	223	—	—	—	
					(020)	(263 58)	(1762 32)	(1848 21)	(2283 95)	(2039 93)			(116)	(122)				
E	808	165	7905 14	—	174*	−730 04*	−1131 32	−1262 90	−2953 11	4002 55**	—	—	—	—	.795**	160	00003*	
					(020)	(261 83)	(1751.53)	(1838 02)	(2262 73)	(2020 89)					(362)	(128)	(00001)	
								Net Outlay on Durables[c] Regressions										
A′	886	.159	80 24	01336*	− 00007	6 26	−15 75	−99 37*	−122 62*	10 27	—	—	—	—	—	—	—	437 20*
				(00199)	(00018)	(5 24)	(35 63)	(36 88)	(44.90)	(40 89)								(52.32)
B′	880	159	69 10	01328*	− 00005	6 15	−18.33	−97 89*	−120 67*	13 79	23.57	—	—	—	—	—	—	435 80*
				(00201)	(.00018)	(5.29)	(35 82)	(37 11)	(45 33)	(41 44)	(26 55)							(52 50)
C′	871	164	65 49	01314*	− 00004	6 99	−14 20	−102 46*	−119 91*	10 68	—	01471	—	—	—	—	—	441.99*
				(00201)	(.00018)	(5 27)	(35 89)	(37 09)	(45 17)	(41 27)		(01607)						(52 63)
D′	880	159	79 81	—	− 00004	6 04	−19 19	−98 26*	−121 04*	14 22	—	—	01447*	01236*	—	—	—	436 45*
					(00018)	(5 29)	(35 82)	(37 06)	(45.28)	(41 37)			(00223)	(00218)				(52 47)
E′	886	162	81 31	—	− 00002	5 70	−17.96	−99 70*	−125 06*	16 71	—	—	—	—	02259*	01229*	—	432.38*
					(00018)	(5 24)	(35 60)	(36 84)	(44 86)	(41 00)					(00552)	(00208)	—	(42.32)

*indicates significances at 1 percent level

**at 5 percent level.

[a]Numbers in parentheses are standard errors

[b]See (1) in text for definition of consumption

[c]Durables are defined as furniture, refrigerators, washing machines, stoves, television sets, household appliances, and air conditioners.

In regressions B' and C' respectively, neither *WDIN68* nor *HRSYR* is significant. Moreover, when regression D' and E' are compared with regression A', the Chow tests indicate that we cannot reject the null hypotheses that the two regressions are structurally the same ($F_{2,800} = .99$; $F_{2,800} = 1.4$). (The interaction term $Y_W \times Y_{OF}$ is excluded from E' because it was found to be insignificant.)

Durables regressions were also run for only those 489 families who made durables purchases in 1968. In these regressions, as expected, the coefficient on Y increased somewhat, to about .018, while the coefficient on *MOVHSREC* fell to about \$360. (Mean net outlay on durables was about \$500 for these families.) The coefficients on *H*45–54 and *H*55–64 also fell and became insignificant. However, there were no changes with respect to the insignificance of wives' work on durables expenditures, once total family income was taken into account.

IV. Conclusion

Wives' earnings tend, on the average, to raise *W-W* family incomes to the level of *N-W-W* family incomes in the same life-cycle group. Total family income held constant, the *Dur/Y* ratio is the same for *W-W* and *N-W-W* families; however, the *C/Y* ratio is higher for *W-W* families. Given the increasing labor force participation *(LFP)* among wives, these findings have important implications for employment, price stability and economic growth. However, space constraints permit only a cursory examination of these implications.

A higher *C/Y* ratio for *W-W* families may help to create employment and/or increase pressures on prices. The effects on employment and prices depend upon exactly which goods and services are demanded and on the relative degree of tightness and monopoly and monopsony power in the relevant goods and labor markets. For example, if highly labor-intensive child care services are demanded and the labor and product markets associated with these services are rather loose and competitive, then at least in the "first round," an increased demand for child care is likely to have substantial employment effects and few price effects. To the extent, on the other hand, that increased automobile-associated goods (e.g., gasoline) are demanded, there may be substantial price effects but few employment effects.

Possible effects on growth are even more complicated to assess. Generally, higher savings rates are associated with a more rapid rate of economic growth. However, productivity and the rate of growth of labor input (population and labor force) also affect economic growth. Before we can begin to measure the overall relationship between wives' employment and economic growth, considerable further investigation is required with respect to the effects of wives' employment on productivity and the birth rate.

REFERENCES

R. Agarwala and J. Drinkwater, "Consumption Functions With Shifting Parameters Due to Socio-Economic Factors," *Rev. Econ. Statist.*, Feb. 1972, *54*, 89–96.

M. S. Carroll, "The Working Wife and Her Family's Economic Position," *Mon. Lab. Rev.*, April 1962, *85*, 366–74.

James S. Duesenberry, *Income, Saving and the Theory of Consumer Behavior*, Cambridge, Mass. 1949.

Milton Friedman, *A Theory of The Consumption Function*, Princeton 1957.

John K. Galbraith, *Economics and The Public Purpose*, Boston 1973.

Robert Holbrook and Frank Stafford, "The Propensity To Consume Separate Types of Income: A Generalized Permanent Income Hypothesis," *Econometrica*, Jan. 1971, *39*, 1–21.

Lucy Mallan, "Financial Patterns in Households with Working Wives," unpublished Ph.D. Dissertation, Northwestern University 1968.

Jacob Mincer, "Employment and Consumption," *Rev. Econ. Statist.*, Feb. 1960, *42*, 20–26.

______, "Labor Supply, Family Income and Consumption," *Amer. Econ. Rev. Proc.*, May 1960, *50*, 574–83.

Joann Vanek, "Time Spent in Housework," *Scientific American*, Nov. 1974, *231*, 116–20.

Capacity: An Integrated Micro and Macro Analysis

By GORDON C. WINSTON*

"In principle, 'capacity' has meaning."
—Alan Greenspan

This paper reports on a longer study that constructs an integrated description of productive capacity. There is too little space here carefully to review the logic of that analysis; instead, this paper will describe what that analysis does and how it does it.

The quotation above nicely conveys the simultaneous sense of faith and frustration that permeates the idea of productive capacity for a firm or an economy. It is an idea that has proved as useful on a macroanalytic level as it has been intractable on the level of the individual firm. The purpose of the study was to describe capacity in a way that makes sense simultaneously at both micro and macroeconomic levels and shows why a firm would generate the familiar behavior with respect to its capacity that we expect from an economy with respect to aggregate capacity.

I. Macro Capacity

In macroeconomics, 'capacity has meaning' in a number of areas of analysis:

> Some idea of capacity and its utilization is central in investment demand analysis; the accelerator is damped by excess capacity since, in its presence, increases in product demand do not induce further demand for investment goods.
>
> Some idea of a capacity ceiling to real output is embodied in analyses of the real causes of price inflation.
>
> Capacity helps explain variations in trade flows—a more pressing concern in *U.K.* studies than in the United States—since excess capacity acts as an inducement to increase exports.
>
> Changes in factor productivity over the business cycle that result in changes in income shares are intimately related to changes in utilization of capacity.

Underlying these applications is a generally accepted idea of aggregate productive capacity that would go something like this: "Capacity is the maximum sustainable level of output (per year) that can be got when an economy's available resources are fully and efficiently employed, given tastes and technology."

All of this is at a macroeconomic level. On the other end, it has been very hard to envision what, precisely, is happening in the typical firm when the economy is operating at capacity or, harder yet, when the economy is operating below capacity.

II. Micro Capacity

While there is little difficulty with our definition of capacity on a macro level, it creates problems at a micro level since it has two aspects that conflict within an individual firm. First, it says that capacity is a maximum output. This implies a technical definition of capacity for the firm; an *engineering capacity*. But at the same time, it says that capacity is a *most efficient* level of output and that, for the individual firm, suggest and *economic capacity*, which is quite different from an engineering or technical maximum output. In economic capacity, costs become important. Lawrence Klein, George Perry and others have therefore suggested that the firm's capacity should be defined as the output that achieves lowest average costs. Frank deLeeuw defined the firm's capacity with respect to its marginal costs and their relation to average costs. Both of these concepts imply that the firm's level of economic capacity is significantly below what technically it could produce as an engineering maximum.

*Williams College

Thus a major part of the problem has been that the macro conception of capacity rests on both maximum and efficient output but that those goals are in conflict within the firm.

Compounding this is a sense frequently encountered in the literature that productive capacity should be a straightforward technical matter like the capacity of a bucket and not something we have to haggle over with economic subtleties. A five-gallon bucket has a capacity of five gallons and a one-hundred-ton-a-day plant should, by the same reasoning, have a capacity of one hundred tons a day. That should be the end of that. If follows that capacity output should be determined solely by the capital stock—by the bucket—both for the firm and for the economy as a whole. But this underlines two further anomolies. What about other resources? In the macro conception of capacity, the concern is certainly with the availability and use of all of the economy's resources and not with capital alone. Second, and perhaps more difficult, how can it be that firms' capital is idle most of the time, if capital stocks define output capacity?

III. An Integrated Conception of Capacity

The tact taken in this analysis is that (1) macro capacity is a general equilibrium concept and (2) in that general equilibrium, resource allocation is efficient when plants' capital stocks are idle much of the time. The first of these is familiar; it has been said frequently by Klein among others. The second is not so familiar. That idle capital is usually optimal is a new insight that depends on a time-specific or optimal utilization analysis of the firm's production decisions; an analysis that is based on Robin Marris' study of capacity utilization in the United Kingdom, on Nicholas Georgescu-Roegens' description of production and on my work in refining, formalizing and extending utilization theory, alone and with Thomas McCoy.

The idea underlying this model of production is simple. Firms know when they invest that some of their input costs are going to vary rhythmically over calendar periods; those costs will always be high during certain regular times of the day or the year and they will always be low at other times of the day or the year. This fact is known, predictable and trendless. Labor, for instance, is typically higher priced at night when firms have to pay a night-shift wage premium to induce people to work at a generally unpleasant hour. Agriculturally based inputs to production have a seasonal pattern so they are typically higher priced before than after harvest. Input price rhythms appear to be ubiquitous.

A firm operating in an environment of rhythmic input prices, if it increases the proportion of time it operates its plant and equipment (the utilization of its capital stock), will achieve lower and lower average *capital* costs but it will have to pay higher and higher prices for the rhythmically priced input. Inauguration of a third daily shift in a manufacturing plant, for instance, will reduce average capital costs but it will increase labor costs, too. The *optimal* utilization for a plant, then, reflects a balance between these two opposing forces, the reduction in capital costs achieved by higher utilization and the increase in the costs of rhythmically priced inputs that accompanies it. That optimal balance will often leave the plant idle a good deal of the time. The most dramatic example I have encountered is a sugar mill in Louisiana that operates thirty-one days a year, when cane is available at low prices, and shuts down the other 91 percent of the time when cane prices are high.

The implication of this analysis is that the most efficient, least cost, level of utilization (and output) for a firm will often be a good deal less than the maximum level of utilization (and output) at which technically it could operate. The typical firm could always produce a good deal more output per year simply by utilizing its capital more of the time but to do so would increase its costs of production. So the firm's least cost level of output, its economic capacity, comes when its capital stock is idle much of the time and economic capacity is therefore a level of output that is less than the engineering or technical maximum.

Unfortunately, noneconomic (maximum or engineering) capacity is sometimes, apparently

unintentionally, transformed into economic capacity when it is defined with the proviso that output is at a "technical" maximum "under standard and normal hours of operation" (as, for instance, by Leif Johansen). But, this can hardly be an engineering concept of capacity when those "standard and normal hours of operation" are, themselves, the result of an economic decision—a fact made clear by optimal utilization models—since the firm's economic capacity output and its standard and normal hours of operation respond to changes in relative prices and price rhythms. Nor is the error entirely innocent. It represents as a technical fact of life what may well be the result of a dubious economic policy—as when poor countries induce low utilization of their capital stocks by their factor price policies or by legislated nighttime wage differentials.

In the integration of the micro and macro manifestations of capacity, a key insight from optimal utilization analysis is that when the economy as a whole is operating at its macro capacity level of output, the individual firms that make up that economy are operating at an economic capacity level of output, therefore with a good deal of idle capital. Idle plants and maximum aggregate output coincide because it is efficient for firms *not* to use their capital all the time. The efficient economy in its turn will have adjusted its resource allocation to reflect that fact and maximum aggregate output will not require (indeed, would be inconsistent with) output from each individual firm.[1] A potential fallacy of composition is involved since in an aggregate general equilibrium at full capacity, any one firm alone typically might increase its output beyond its economic capacity—even to its engineering capacity—but all firms considered together could not, simply because in the aggregate there are not enough available resources. Resources will have been allocated efficiently so that firms are operating at their economic capacities with—because they *are* operating efficiently—a good deal of idle capital.

IV. The Significance of an Integrated Capacity Analysis

What is the significance of this integrated model of capacity? First, and most simply, it is neat; it is satisfying to be able finally to describe productive capacity in a way that allows us to see what is happening in a rationally managed firm and how that fits consistently into what is happening in the economy as a whole. An area of uncomfortable ambiguity is eliminated.

Second, in this model, profit maximizing firms behave with respect to their capacity exactly the same way we expect an economy to behave with respect to its aggregate capacity:

> A firm will operate at a level of output above capacity if it is paid a high enough price for its product to cover the increased marginal costs. So operation at outputs in excess of capacity levels is, indeed, to be expected under some conditions, but it is, indeed, also inflationary
>
> Sustained operation in excess of capacity will induce the firm to expand its capital stock by investment in order to reduce utilization back to its least cost level.[2]

[1]The more detailed argument expands on this, noting that an economy's "maximum *available* resources"—which define maximum aggregate output—will always be less than the maximum *possible* flow of resources that *could* be got from its factor stocks and its population This distinction is clearest for labor where we typically accept, as a different though economic question, the factor owners' (households') decisions on how much resource flow (labor service) they choose to make available from their factor stocks (population) and how much they choose to withhold (take as leisure) War and other passions may induce a change in work preferences that would change the available labor resources from a given population, hence, aggregate capacity. But that is beside the present point.

[2]I recently conducted a survey of forty-five firms in Nigeria, an economy with a very high level of aggregate demand As predicted by this model of capacity, a number of firms reported that they were operating well beyond their capacity levels—some at as much as 150 percent They said they were doing so because product prices were high but they said, too, that they had explicit plans for investment that would allow them to reduce utilization back to the levels they considered to be least cost—back to economic capacity operation with more capital stock and the larger output

Whether one firm alone or all firms together operate above their capacity levels is important because that will determine how steeply their marginal costs will rise after economic capacity levels are reached.

Exports are encouraged by excess economic capacity as firms seek to expand output.

Cyclical labor (and capital) productivity changes will result if firms are reluctant to adjust their labor input as output varies around economic capacity levels.

In all these respects, firms' microeconomic behavior parallels what we have come to expect for the economy as a whole.

Third, this analysis emphasizes that capacity should be defined, at both micro and macro levels, with respect to all resources and not with respect simply to capital stock. This supports the warnings (by Perry, Klein, *inter alia*) that overestimation of capacities will result when individual sectors are considered in isolation from the rest of the economy.

Fourth, this analysis shows that the empirical capacity utilization measures typically generated in the United States simply measure different things. The McGraw-Hill capacity utilization series is widely thought to measure firms' current operations relative to their most economical, least-cost levels of output, in our terms, McGraw-Hill is an estimate of firm's economic capacity utilization. The Wharton utilization series, in contrast, takes as capacity observed past peak output (at the two-digit level, then aggregated) and expresses current output as a percent of that; in our terms, Wharton capacity is based on firms' marginal cost. Wharton and McGraw-Hill measures, therefore, would frequently differ, both in magnitude and direction, depending on the data served up by recent history.[3] The McGraw-Hill estimates of economic capacity are thought to be influenced by animal spirits—by optimism or pessimism based on current output. They measure in principle something quite solidly defined at the level of the firm. The Wharton series, in contrast, reports on movements of output up and down firms' marginal cost curves relative to movements of peaks up and down firms' marginal cost curves, while the shape of those curves is being changed by the proportion of the economy that is trying to move in the same direction and the same time. Even with constant capacity, different patterns of fluctuation in output would generate quite different Wharton utilization rates. So it seems inevitable that the vagaries of history would have generated disparities between these two indices in both trend and level, disparities that only an integrated model might hope to sort out.

Finally in this analysis, there are seen to be three kinds of social excess capacity, two of which are familiar, one of which is not. The most familiar is the excess capacity that occurs when firms do not produce up to their targets of economic capacity output, a problem that results from Keynesian deficient demand (typical in advanced countries) or from deficient supplies of inputs (typical in less developed countries)—both, of course, at prevailing prices. A different, but still familiar, sort of excess capacity appears when firms face scale economies with limited product markets and therefore operate a "too-large" plant at less than its economic capacity because that minimizes cost—in textbook graphics, they operate at a Viner tangency of falling long- and short-run average costs. The third and unfamiliar kind of social excess capacity appears when firms set their economic capacity targets too low in light of society's real scarcities; they may report "full utilization of capacity," but what the firms see as "full" is not full enough because they base their private calculations of economic capacity on input prices that do not reflect real scarcities—the integrated capacity analysis shows that artificially cheap capital, for an important instance, would induce firms to set full capacity targets lower than they should.

[3]The third and "most eclectic of the indexes" of capacity (Perry, p. 707)—the Federal Reserve's—is an amalgam of the McGraw-Hill index, capital stock estimates and assumptions about optimal capital productivity that defies simple classification in a model of the firm.

The analysis reported on in this paper thus provides a compatible microfoundation for the useful concept of aggregate productive capacity. Aside from the satisfaction of consistency per se, this integrated analysis has promising implications for understanding the behavior of empirical measures of capacity, for recognizing the role of economic variables in defining full capacity output and for identifying the sources of social excess capacity.

REFERENCES

Frank de Leeuw, "The Concept of Capacity," *J. Amer. Statist. Assn.*, 1962, 320–29.

Nicholas Georgescu-Roegen, "Chamberlin's New Economics and the Unit of Production," ch. 2 in R. E Kuenne, ed., *Monopolistic Competition Theory: Studies in Impact*, New York 1967, 31–62.

Leif Johansen, "Production Functions and the Concept of Capacity," *Recherches recentes sure la Fonction de Production*, Collection "Economie mathematique et econometrie," No. 2, Ceruna, Namur 1968.

Lawrence R. Klein, "Some Theoretical Issues in the Measurement of Capacity," *Econometrica*, Apr. 1960, *28*, 272–86.

——— **and Virginia Long,** "Capacity Utilization: Concept, Measurement and Recent Estimates," *Brookings Papers*, 1973, *3*, 743–56.

Robin Marris, *The Economics of Capital Utilization: A Report on Multiple-Shift Work*, Cambridge 1964.

George L. Perry, "Capacity in Manufacturing," *Brookings Papers*, 1973, *3*, 701–42.

Gordon C. Winston, "The Theory of Capital Utilization and Idleness," *J. Econ. Lit.*, Dec. 1974.

——— **and Thomas O. McCoy,** "Investment and the Optimal Idleness of Capital," *Rev. Econ. Stud.*, July 1974.

———, "The Concept of Production Capacity: An Integrated Micro and Macro Analysis," Williams College, June 1976.

A General Equilibrium Approach to Estimating the Costs of Domestic Distortions

By JAIME A P. DE MELO*

Recent developments in the field of trade policy have been dominated by the elaboration of the theory of domestic distortions in open economies. Essentially the new approach focuses on the choice between alternative policies and provides a greatly improved method of analyzing the effects of trade policies. (See articles by Jagdish Bhagwati and Stephen Magee, 1973.) The analytical rigor and intellectual effort which has gone into this theoretical work has not been matched on the empirical side, where the measurement of the welfare costs of domestic distortions has concentrated on trade distortions and has most often been carried out in a partial equilibrium framework. The main objective of this paper is to examine the context within which the static welfare costs of distortions have usually been estimated and to outline an alternative approach to measurement based on a general equilibrium analysis. Some fixed point estimates based on Colombian data are provided to illustrate this approach. The framework used in measuring the welfare costs is thus in closer agreement with the theoretical analysis.

I. The Cost of Distortions in Domestic Markets

Domestic distortions due to government intervention or market imperfections create a wedge between domestic prices and domestic opportunity costs. Abstracting from dynamic effects on capital accumulation and growth, and from the possibility of monopolistic power in world markets, the first order conditions for a welfare maximum require that for any pair of commodities the marginal rate of transformation in domestic production equal the foreign marginal rate of transformation and the marginal rate of substitution in consumption ($DRT = FRT = DRS$). The general equilibrium theory of domestic distortions has both analyzed in depth the efficiency implications of a breakdown of these marginal equivalences and provided us with a ranking of policies—in terms of their effect on welfare—to correct these distortions.

On the empirical side, practically all attempts at measuring the costs of domestic distortions have been based upon Harry Johnson's (1960) revival of consumer's and producer's surplus measures of the gains from trade. These studies have usually concentrated on estimating the costs of protection. Based on Hicksian compensation tests and calculus techniques, they have derived expressions for the total cost of protection in terms of elasticities of compensated demand and supply. This approach is confined to small departures from free trade.

Models developed for the actual purpose of measurement have gone through four stages. In the first stage (see W. M. Corden, 1957), models were developed at an aggregate level in which only total imports and total exports were considered. The second stage (see Johnson, 1958), saw the disaggregation of the analysis to include the effects due to the relationships of substitution and complementarity among importable goods. In the following stage (see Giorgio Basevi, Magee 1972), the analysis was extended

*Assistant Professor, Georgetown University. My appreciation to Bela Balassa for stimulating my interests in this area and to Michael Crosswell and Peter Kenen for helpful comments The research upon which this paper is based was carried out while I was with the U.S Agency for International Development The views expressed here are mine and are not intended as statements of Agency policy Remaining errors are my responsibility.

to include the terms of trade effect and some general equilibrium aspects with the concept of the "uniform tariff equivalent" for the purpose of calculating an index of tariff structures. In the fourth stage (see Bela Balassa) the process of disaggregation was extended to include intermediate products, and nominal rates of protection were replaced by effective rates (*ERP*s) in the calculation of the cost of protection. Furthermore some steps were taken to classify goods according to whether their production would cease under free trade.[1]

The models developed for the actual measurement of protection have reached an impressive degree of sophistication. In light of the available data and shortcomings of the existing parameter estimates upon which any empirical work must rely, they may represent the best compromise between rigor and relevance.

Yet these models suffer from several shortcomings. First, and most important, practically all estimates of distortions have centered around the cost of distortions in commodity markets. Distortions in capital and labor markets (see Ronald McKinnon, Magee 1973) and their impact on international trade have played an increasing role in the theory of domestic distortions; and in providing guidelines for reforming the system of incentives in developing countries, economists have urged that these distortions be removed wherever possible. Moreover, the welfare effect of distortions in commodity markets should not be studied in isolation from existing distortions in factor markets since it is well-known from the theory of domestic distortions that, even in the absence of terms of trade effects, removing distortions in the commodity markets in the presence of distortions in factor markets may lead to a decrease in welfare. Few estimates have been made of the loss in efficiency resulting from the presence of these distortions in factor markets. In a well-known paper Arnold Harberger estimated the costs of distortions in factor markets in Chile. More recently estimates by Christopher Dougherty and Marcelo Selowsky for Colombia and by Gunnar Floystad for Norway have found these efficiency losses to be small.

Second, there is the difficulty of marrying partial equilibrium estimates with general equilibrium analysis, which has been discussed at length in the literature on the existence of an *ERP* index of resource tariffs. The defects of Effective Protection as a general equilibrium theory—viz. the failure to admit substitution possibilities among inputs, factor price flexibility and the treatment of nontraded goods whose prices are determined in domestic markets—carry over to the estimates of the costs of protection.[2] Thus it is not clear how one goes about estimating general equilibrium demand and supply curves

Some attempts have been made at quantifying the effects of distortions within a general equilibrium framework. They have concentrated on trade distortions, and have taken place in a programming framework that optimizes an explicit objective function (see David Evans). This approach, however, is not particularly well suited for the purpose at hand since there are well-known difficulties in analyzing the effects of changing pre-existing price wedges such as tariffs, taxes, and distortions in factor markets. Within this approach the interpretation of a tariff-ridden balance of payments can be difficult (see Lance Taylor). Finally the incorporation of substitution in demand and supply —essential in the face of large changes in relative prices among factors and commodities —is computationally cumbersome.

[1]Concurrently Joel Bergsman adapted the concept of "*X*-efficiency" to reflect the loss in efficiency and productivity resulting from protection While this phenomenon may indeed be important especially in less developed countries, it is not an integral part of the theory of domestic distortions and is beyond the purview of this paper

[2]In a recent paper on welfare surpluses and the gains from trade, James Anderson notes that Johnson's expanded method is far from our reach as a method of measurement despite its theoretical appeal (p. 762) See John Whalley for a discussion of the inherent difficulties in comparing estimates derived from partial and general equilibrium analyses

The following section outlines an approach based on general equilibrium analysis which seems well suited for studying the global effects of distortions in domestic commodity and factor markets.

II. General Equilibrium Estimates of the Costs of Market Distortions

A. *The Walrasian Approach*

An alternative approach is to start from a Walrasian description of the economy where optimization is implicitly carried out separately by consumers and producers who maximize utility and profits subject to budget and production constraints. A distinctive feature of this approach is that the various agents in the economy may interact through a variety of specifications of market behavior which are particularly well suited for the incorporation of price distortions in commodity and factor markets.[3] Essentially the solution of the problem is that of finding a fixed point for a set of simultaneous nonlinear demand and supply equations.

This approach should be useful in future attempts at quantifying the costs of market distortions, particularly in developing countries where simulation of market behavior requires careful portraying of a large variety of economic and institutional rules However, both the quality of the concomitant data requirements and the difficulty of ascertaining genuine market distortions from divergences which may be the result of nonmaximizing behavior should be kept in mind when interpreting results derived from this approach. With these caveats in mind we turn to some illustrative estimates obtained from an implementation with Colombian data.

B. *Illustrative Estimates for Colombia*

Some salient characteristics of the model are briefly outlined here and the reader is referred to de Melo for a full presentation of the model and a discussion of data sources. On the supply side, producers maximize profits subject to a Leontief technology for intermediate inputs and noncompetitive imports. For value-added, we specify Cobb-Douglas and two-level constant elasticity of substitution production functions. All primary factors (land, capital, skilled and unskilled labor) are in inelastic supply and are fully employed; factor returns are endogenously determined. Distortions in factor markets are introduced by specifying a differential between the price of an identical factor in different sectors. Several assumptions are made about intersectoral capital mobility and the migration of factors between rural and urban sectors is limited to emphasize the dichotomy between these segments of the economy. On the demand side the representative consumer maximizes a Stone-Geary utility function which generates the linear expenditure system. Distortions in the commodity markets arise from tariffs, taxes, and subsidies on traded goods. Finally, in line with recent developments in the pure theory of trade, a distinction is made between tradable goods whose prices are determined in world markets with quantities traded clearing the domestic markets, and home goods whose prices adjust to clear their markets. The model includes fifteen sectors, of which four are classified as nontraded, and the small country assumption is maintained for traded goods with the exception of coffee which faces a quota on its exports. The exchange rate is endogenously determined so as to maintain equilibrium in the balance of payments.

TABLE 1—DESCRIPTION OF EXPERIMENTS

Experiment	
A–1	Remove tariffs and subsidies
A–2	Same as A–1, but fix sectoral capital stocks
B–1	Remove distortions in labor markets
B–2	Same as B–1, but fix sectoral capital stocks
B–3	Remove distortions in capital markets
B–4	Remove distortions in all factor markets
C–1	Combine A–1 and B–4
C–2	Combine A–1 and B–2

[3]See Leif Johansen for the formulation of the first Walrasian model Irma Adelman and Sherman Robinson provide an example of the variety of market clearing principles which can be captured by this approach.

TABLE 2—COSTS OF DISTORTIONS IN DOMESTIC MARKETS
(Percent change from initial situation)

	Experiments							
	A-1	A-2	B-1	B-2	B-3	B-4	C-1	C-2
Welfare[a]	.6	.0	3 1	1 7	3 7	5 3	4.0	2.1
GNP at Current Domestic Prices	-4.5	-3.6	37.8	8.8	28 0	45 0	12 0	4 9
Exchange Rate Adjustment[b]	6 9	10 7	-15.0	-9 5	-11 4	-22 4	0	1 0

[a]Measured by the *Utility Indicator*
[b]Devaluation (+), Revaluation (−)

Some experiments are listed in Table 1 with corresponding estimates of the costs of distortions reported in Table 2. Removing tariffs and subsidies consists of setting domestic price equal to world price in all traded sectors with the exception of the coffee sector where the export tax is adjusted so that, after meeting domestic demand, producers supply a predetermined quantity of exports equal to Colombia's share in the international coffee agreement. In factor markets, credit rationing generally favoring import-substituting investments along with minimum wage legislation and trade union pressure in the manufacturing sector are taken to be represented by different rates of return to capital and labor across urban sectors. Removing distortions in factor markets then consists of equalizing factor returns across urban sectors. In line with migration theory, the flow of factors between rural and urban sectors is limited by maintaining a fixed differential between rural and urban factor wages.

The results in Table 2 indicate the importance of capital mobility in determining the costs of distortions in factor and commodity markets. It can be seen that the efficiency losses from distortions in factor markets are substantial. A comparison of experiments B-4 and C-1 illustrates the importance of viewing distortions together. One would suspect that removing distortions in both factor and commodity markets (C-1) would lead to greater welfare gains than removing distortions in factor markets alone (B-4). This, however, is not true because in the former case the economy is further away from its optimal quota than in the latter and the resulting losses outweigh the gains from removing distortions in commodity markets. This counterintuitive result can be better understood by noting that raising the export tax on coffee (initially set at 44 percent) to maintain the quota share on world markets lowers welfare since it amounts to impeding specialization. In experiment B-4, the export tax is lowered by 46 percent from its initial value, while in experiment C-1 it is raised by 6 percent. This adjustment in the coffee tax rate also accounts for the substantial difference in exchange rate adjustments between these two experiments.[4]

The importance of general equilibrium effects is illustrated by the experiments removing distortions in factor markets which show substantial revaluation as the value of nontraded goods is raised following an increase in factor wages in these sectors.[5] Although the large efficiency

[4]This type of result raises the question about the optimal level of intervention in domestic markets in the face of unremovable distortions such as quotas and rigidities in some factor markets. Addressing this important issue should preferably take place in a dynamic framework since, for policy purposes, it is desirable to consider simultaneously the static and dynamic effects of interventions in these markets. From a practical point of view such a task is an ambitious one since it amounts to optimizing utility subject to a large system of simultaneous nonlinear equations with a large number of policy instruments.

[5]*GNP* is usually measured at world prices since it serves as a proxy for indicating the alternatives open to the economy. Here the cost of distortions is measured directly by the change in the utility indicator. Moreover, relative prices of nontraded goods are affected by distortions in the price of traded goods. Since they cannot be translated into world prices, only *GNP* expressed in domestic prices has been reported in the results.

gains registered in the table are a function of the parameters entering the model, it appears that the low magnitude of the efficiency gains indicated by previous estimates were due in part to the methodology and assumptions in these studies resulting in both fixed product and factor prices and limited substitution possibilities in consumption and production.

III. Conclusion

The Walrasian approach offers a great number of possibilities to assess the efficiency implications of alternative policies in a general equilibrium framework. The use of multisector economy-wide models provides a better understanding of the impact of policy changes on the structure of the economy.

However, as tools for policy implementation, multisector trade models are prone toward exhibiting extreme specialization because traded sectors whose domestic prices are generally assumed to be fixed to world prices outnumber primary factors of production whose prices are free to vary. Both theoretical and empirical work beyond dropping the small country assumption needs to be done to overcome these strong specialization tendencies. For example, allowing for imperfect substitution between foreign and domestic goods and exploring the possibilities of modeling two-way trade deserve further attention.

Beyond extensions in the specification of foreign trade in multisector models, estimation of the costs of distortions in domestic markets needs to be undertaken in a dynamic framework. The benefits of protecting sectors with rapid technical progress or greater "learning" potential may then be evaluated against the static welfare losses resulting from a misallocation of resources. Similar comparisons may be made of the costs and benefits of credit rationing, minimum wages, etc., in various sectors. Then the challenge in estimating the costs of market imperfections and interventions is twofold: 1) understanding the nature of equilibria in different markets in a dynamic sense (adjustments to disequilibrium are likely to be more rapid in product than factor markets), and 2) successfully distinguishing situations which are a cause of distortions from those which are endogenous to the economic system and are not amenable to policy intervention.

REFERENCES

Irma Adelman and Sherman Robinson, *Income Distribution Policy in Developing Countries: A Case Study of Korea*, Stanford 1976, forthcoming.

James Anderson, "A Note on Welfare Surpluses and Gains from Trade in General Equilibrium," *Amer. Econ. Rev.*, Sept. 1974, *64*, 758–62.

Bela Balassa, *The Structure of Protection in Developing Countries*, Baltimore 1971.

Giorgio Basevi, "The Restrictive Effect of the U.S. Tariff," *Amer. Econ. Rev.*, Sept. 1968, *58*, 840–52

Joel Bergsman, "Commercial Policy, Allocative Efficiency, and 'X-Efficiency'," *Quart. J. Econ.*, Aug. 1974, *88*, 409–433.

Jagdish Bhagwati, "The Generalized Theory of Distortions and Welfare" in J. Bhagwati, et al (eds.), *Trade, Balance of Payments and Growth: Papers in Honor of C. Kindleberger*, Amsterdam 1971.

W. M. Corden, "The Calculation of the Cost of Protection," *Economic Rec.*, April 1957, *33*, 29–51.

______, *Trade Policy and Economic Welfare*, Oxford 1974.

Jaime A. P. de Melo, "A Multi-Sector, Price Endogenous Trade Model Applied to Colombia," Ph.D. dissertation, The Johns Hopkins University, 1975.

Kemal Dervis, "Substitution, Employment and Intertemporal Equilibrium in a Non-Linear Multi-Sector Planning Model for Turkey," *Eur. Econ. Rev.*, Jan. 1975, *6*, 77–96.

Christopher Dougherty and Marcelo Selowsky, "Measuring the Effects of the Misallocation of Labor," *Rev. Econ. Statist.* Sept. 1972, *54*, 386–90.

H. David Evans, "Effects of Protection in a General Equilibrium Framework," *Rev. Econ. Statist.*, May 1971, *53*, 147–56.

Gunner Floystad, "Distortions in the Factor Market: An Empirical Investigation", *Rev.*

Econ. Statist., May 1975, *57*, 200–13.

Arnold Harberger, "Using the Resources at Hand More Effectively", *Amer. Econ. Rev. Proc.*, May 1959, *49*, 134–46.

Leif Johansen, *A Multi-Sectoral Study of Economic Growth*, Amsterdam 1960.

Harry Johnson, "The Gains from Freer Trade with Europe: An Estimate," *Man. School Econ. Soc. Stud.*, Sept. 1958, *26*, 247–55.

———, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, *68*, 327–45.

Stephen Magee, "Factor Market Distortions, Production and Trade: A Survey," *Oxford Econ. Papers*, Nov. 1973.

———, "The Welfare Effects of Restriction on U.S. Trade," *Brookings Paper*, *3*, 1972, 645–701.

Ronald McKinnon, *Money and Capital in Economic Development*, Washington 1973.

Lance Taylor, "Theoretical Foundations and Technical Implications", in C. Blitzer, et al (eds.) *Economy-Wide Models and Development Planning*, 33–110, Oxford 1975.

John Whalley, "How Reliable is Partial Equilibrium Analysis?" *Rev. Econ. Statist.*, Aug. 1975, *57*, 299–310.

Agricultural Development on the Frontier: The Case of Siberia Under Nicholas II

By DANIEL R. KAZMER*

This paper presents highlights of a theory of the economic interactions which determined the nature and extent of Siberian agricultural development after the opening, in 1896, of the part of the Trans-Siberian Railway that connected Western Siberia with European Russia. The history of this development has been thoroughly discussed elsewhere (D. W. Treadgold and Kazmer).

I. Land Use on the Frontier

Most economists are familiar with the Lewis model of development with unlimited supplies of labor. Agricultural development on the frontier may be viewed analogously as development with unlimited supplies of land. Land use in a competitive agricultural sector with free movement of labor and capital is determined as in Figure 1. The essential element is that, in general, land use cost (*LUC*) rises as more land is brought into use. Thus, *LUC* will generally slope upward The marginal product of land is *MPL*. The total land in use is *OA* while the land use cost of each plot is *OB*. The total land use cost is the area *OADB*. The area *EBD* is the pure rent collected by nonmarginal land. The shaded area *BCD* is the return to labor and capital.

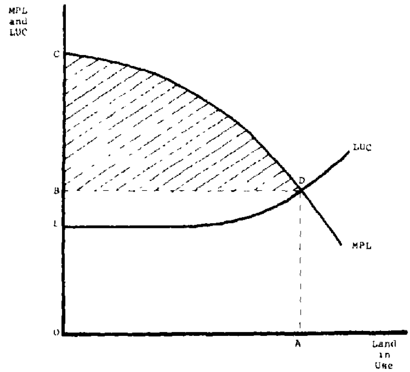


FIGURE 1

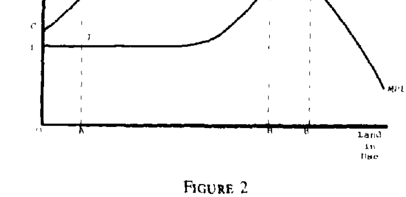


FIGURE 2

Figure 2 explains the timing and extent of the great migration of European Russian peasants into Siberia. The *MPL* curve depicts the marginal product of land in both Siberia and European Russia. The *LUCS* curve represents land use cost in Siberia before the completion of the Western Siberian section of the Trans-Siberian Railway in 1896. The cost is low for well-located, high quality land with amenity benefits for the choicest parcels, which are occupied first. However, as more parcels are brought into use, land use cost rises sharply due

*Central Intelligence Agency. Many thanks to the members of my thesis committee—Evsey Domar, Peter Temin, and Charles Kindleberger—who commented on more primitive versions of this model The errors which lurk within these pages are, of course, my own. A complete paper is available from the author.

to rapidly rising transport, migration, and homestead start-up costs at the margin. *LUCER* represents land use costs in European Russia both before and after 1896. European Russia, with its relatively more developed transport network has less steeply rising land use costs. Since peasants tend to move toward parcels with the lowest land use costs, *OA* represents the quantity of land in use in Siberia before 1896, and *AB* represents the land in use in European Russia. The area *CDE* represents the reduction in land use cost gained by moving to Siberia.

With the completion of the railroad link in 1896, however, a new curve must be added. *LUCSR* represents the new land use cost curve in Siberia with rail transport. It is horizontal for most of its length and then rises sharply. The horizontal portion refers to land close to the railroad, which had a well-developed transport network linked to the railroad. Farther from the railroad, however, the costs of transport to the rail station increased rapidly (the steeply rising portion of the *LUCSR* curve) until they became prohibitive at a distance of about 100 miles. With rail service, the quantity of land which will eventually be occupied in Siberia increases from *OA* to *OH* while the land in use in European Russia declines from *AB* to *HB*. Thus, the migration of labor and capital to be expected is that required to work the quantity of land *AH*. The use cost of Siberian land occupied before the construction of the railroad is further reduced by the extra amount *FCEJ* because of the railroad. The area *JEG* represents the reduction in land use cost gained by moving to Siberia after the railway opened.

Figure 3 depicts the dynamics of agricultural expansion on the frontier. Expansion of land use in Siberia will not go beyond the point at which *LUCSR* and *LUCER* intersect; at that point, marginal land use cost in European Russia becomes lower than in Siberia. MPL_1 represents the marginal product of land given some quantity of labor and capital. *OA* units of land will be used to produce total product *OABCD* of which *OABD* is the return to land with no pure rent and *DBC* is the return to labor and capital. Part of this return to labor and capital will be used to augment the supplies of those same factors, and part will be used to set up and expand homesteads thus bringing more land into use. With more labor and capital available, *MPL* shifts up and to the right to, say, MPL_2. Now *OA'* units of land will produce *OA'B'C'D* units of total product of which *OA'B'D* offsets land cost with no pure rent and *DB'C'* is the return to labor and capital. Part of this return is used to expand the supply of labor and capital and bring more land into use, so *MPL* shifts up and to the right again and the process may continue in a similar fashion until *LUCSR* curves upward. Additions to labor and capital shift *MPL* out to, say, MPL_n. The total product is *OEFI* of which only *GFI* now goes to capital and labor and *OEFHD* offsets total land use cost. *DJFG* would have gone to labor and capital, but since *LUCSR* is rising, *HFJ* must be used to offset land use cost and *DHFG* is a pure rent to land which carries a use cost less than that of land at the margin. Since only *GFI* goes to capital and labor; one might expect expansion of the supply of those factors to occur more slowly; furthermore, continued increases in capital and labor and shifts of *MPL* to the right result in higher pure rent for land parcels already in use.

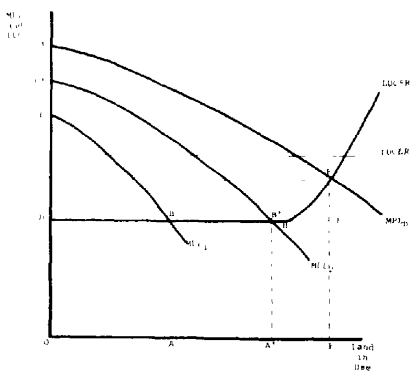


FIGURE 3

This relative distribution of frontier income in favor of nonland factor inputs is an important feature distinguishing frontier from fully-settled agricultural districts. Returns to land and nonland factors go into different pockets—those of new arrivals versus those of established households. New settlers entering a frontier region often do not bring with them the funds needed to cover the homestead start-up and land use costs of even marginal no-rent land. The relative distribution of factor income on the frontier in favor of nonland factors, especially labor, tends to transfer income away from landowning settled peasants toward the new arrivals trying to accumulate the amount necessary to start up a homestead.

II. Labor

The essential features determining the operation of agricultural labor markets were its frontier nature and the role played by immigrants into the region. Figure 4 illustrates this role. A priori reasoning alone can tell us nothing about the shape of the demand curve in Figure 4 However, the short-run supply curve may be expected to be highly inelastic for two reasons. 1) Communication between local markets was poor simply because of distance. 2) Within any local market, the presence of new immigrants favored a more inelastic short-run labor supply curve. New immigrants had to hire themselves out almost regardless of the wage rate to live—but only until they had accumulated the stake necessary to obtain sufficient supplies and equipment to become independent farmers.

A one-shot influx of immigrants into a local labor market shifted the supply and demand curves for hired agricultural labor according to the following pattern (see Figure 4): (a) New immigrant families move into the area. They must work as hired agricultural laborers until they have acquired sufficient capital to become independent farmers. This shifts the labor supply curve to the right. (b) As these same immigrant households amass the necessary capital to become independent farmers, they withdraw from the market for hired agricultural labor to

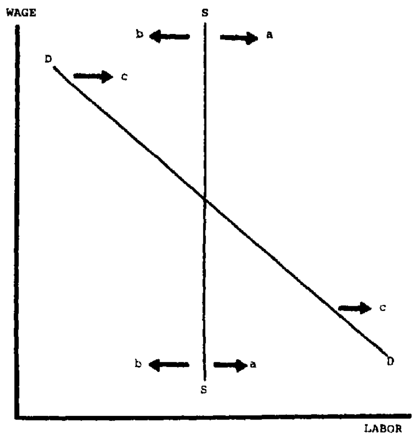


FIGURE 4

work their own plots. This shifts the labor supply curve to the left. The curve will not shift back to its original position since not all immigrant households will be successful in becoming independent farming households. Some will remain in the labor market. This shift to the left occurred approximately seven years after settlement in the Siberian case. (c) As the immigrant household continues to prosper, its need for agricultural labor eventually outstrips its own labor capacity; thus, it must re-enter the locai agricultural labor market as a demander of labor. This shifts the demand curve to the right. This process took place approximately ten years after settlement. Since not all immigrant households which eventually employed all their own members had to seek additional workers on the local market, the rightward shift of the demand curve will be less than the leftward shift of the supply curve described in (b) above and, therefore, also less than the rightward shift of the supply curve described in (a) above. Note that there is no mechanism which guarantees equilibrium in the short run since the supply curve is jolted back and forth by arrivals of new settlers and their withdrawals from the labor market as they accumulate sufficient capital to farm in-

dependently. Nor is there any reason to expect wage rates in different local markets to be similar. This means that no neat conclusions can be drawn from this model. It is, however, an accurate portrayal of the mechanism behind the operations of local markets for hired agricultural labor in Siberia and may shed some light on the operation of local frontier labor markets in other instances.

III. Capital

Agricultural capital may be classified into labor complements and labor substitutes: labor complements increase or decrease in the same direction as the quantity of labor; the quantity of services demanded from labor substitutes changes inversely to the demand for labor services, *ceteris paribus*. For example, if one has a strong son capable of operating a plow, one does not sell one's horse and hitch up one's son; one buys another horse and plot for the son to operate. Hence, the horse and plow act as labor complements; they allow the son's labor power to be used efficiently—the son and horse are not substitutes. On the other hand, a reaping and binding machine does displace labor since it can perform exactly the same function as a harvesting crew. It is a labor substitute.

IV. Conclusions

Several conclusions may be drawn from this analysis. First, there is a definite gain from rising land values to people who purchase or claim choice parcels in the early period of settlement of a frontier area.

Second, turning to local markets for hired agricultural labor, one should not be surprised to find large wage differences from one district or village to another and apparent disequilibrium in frontier labor markets—especially if one must reside on one's land to retain ownership Apparent labor surpluses or shortages on the frontier may be interpreted differently than in a settled region. Unemployment or low wages may indicate only that new settlers arrive to claim land faster than they can be absorbed into employment. On the other hand, labor shortages may be an indicator of the ease of establishing an independent farming household soon after moving in and withdrawing from the labor market—not of the undesirability of an area in general. Perversely, higher wages, if they speed up the process of acquiring the capital needed to become independent, may reduce the supply of labor. The long-term wage level is crucial in determining whether households eventually begin farming independently or are locked into the agricultural proletariat.

Finally, efforts should be made to promote capital imports into a region. However, the distinction between labor complements and labor substitutes is important. Labor complements tend to raise the wage level and encourage more rapid establishment of independent farming households. Labor substitutes, to the extent that they lower the local long-term wage level and raise the start-up costs of an independent farm, may (along with raising output per unit input) have the undesired effect of lengthening the period over which a household must sell its labor on the market before it can become independent. At worst, the household may find itself permanently locked into the proletariat. Might this mechanism be helpful in explaining why the successful frontier development patterns of the 19th century have not been repeated in the underdeveloped countries of the 20th century?

REFERENCES

Daniel R. Kazmer, *The Agricultural Development of Siberia: 1890–1917*, unpublished M.I.T. dissertation 1973.

William A. Lewis, "Economic Development With Unlimited Supplies of Labour," *The Manchester School of Economic and Social Studies*, May 1954, 22, 139–191.

D. W. Treadgold, *The Great Siberian Migration*, Princeton 1957.

PROCEEDINGS

OF THE

EIGHTY-NINTH

ANNUAL

MEETING

ATLANTIC CITY, NEW JERSEY

SEPTEMBER 16-18, 1976

Minutes of the Annual Meeting Atlantic City, New Jersey September 17, 1976

The Eighty-ninth Annual Meeting of the American Economic Association was called to order by President Franco Modigliani in the Pennsylvania Room of Haddon Hall Hotel, Atlantic City, New Jersey at 9:45 p.m. on Friday, September 17, 1976. The President asked that members occupy the first ten rows of seats. He also requested that members identify themselves by name when they spoke from the floor.

The Secretary reported that the minutes of the December 29, 1975 Annual Meeting needed correcting. The second paragraph of Article 1, Section 2 as amended was omitted from the minutes as recorded on pages 465–67 of the *American Economic Review, Papers and Proceedings*, May 1976. The omitted paragraph reads:

> The foregoing dues schedule including income brackets may be increased by the Executive Committee in proportion to the increase occurring after January 1, 1976, in relevant price and wage indexes, provided that the increase in any year shall not exceed ten percent

The quoted paragraph was part of an amendment to Article I, Section 2 of the bylaws approved by the members in 1975 It should be added to page 466, column 1, immediately following the indented paragraph on pages 465–66.

The minutes were approved as corrected.

The Secretary, Treasurer (Rendigs Fels), the Managing Editor of the *American Economic Review* (George H. Borts), the Managing Editor of the *Journal of Economic Literature* (Mark Perlman), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were available at registration and were also distributed at the meeting itself. (See their reports published in this issue.)

The Secretary presented the following resolutions, which were adopted unanimously:

> BE IT RESOLVED that this meeting record a special vote of thanks to the members of the 1976 Allied Social Sciences Associations' Local Arrangements Committee chaired by Edward G. Boehne for their hard work and efficient management of these meetings.

> Be it resolved that this meeting commend Lawrence R. Klein for planning a program of great interest and high distinction.

At this point, President Modigliani introduced President-elect Lawrence Klein, who took the chair. In calling for new business the President-elect stated that four resolutions by members had been submitted to the Secretary a month in advance of the meeting required by the Association's regulations. These resolutions had been distributed before and at the meeting.

The President-elect called for discussion of the resolution submitted by Julian M. Greene (proposer) and Bernard W. Tenenbaum (seconder). Since no one spoke in behalf of the resolution, it was moved, seconded and PASSED that discussion of the resolution be postponed indefinitely.

Two resolutions had been proposed by Alfred Kraessel and seconded by William J. Doerflinger, which read:

> BE IT RESOLVED that the Association accept motions presented on the floor of any business meeting, thus reversing an executive decision adopted earlier this year, which requires that motions be presented in writing one month prior to the business meeting The purpose of this is to bring about increasing participation and democratization of the procedures of the American Economic Association

> That the American Economic Association set up a committee to study the economic status of the profession and the possibility of defining professional economists as different from general practitioners in the fields of business, social sciences, etc

Kraessel spoke in favor of the first resolution.

He stated that allowing members to present motions from the floor would lead to greater participation in the democratization of the Association's affairs. William Vickrey spoke against the motion, saying that since only a minute percentage of members attends the annual business meeting, "surprise" resolutions would be unwelcome. The resolution was put to a vote and lost.

On behalf of his second resolution, Kraessel pointed out that other professionals, such as chemists, accountants, attorneys, physicians, etc., are often afforded protection by their associations and that, in these times, professional economists need such protection by their association. He emphasized the importance of power groups in attracting students, accrediting programs, and maintaining the relative position of economics in the professions. The resolution was put to a vote and failed.

Before calling for discussion of the resolution submitted by Robert Cherry and David M. Gordon, the President-elect stated that Counsel had advised that since the resolution called for a change in the Certification of Incorporation and that since required legal procedures had not been followed, the resolution was out of order and could not be acted upon at this meeting. It was included on the agenda for discussion purposes only. The resolution reads:

> WHEREAS the standing practice of the American Economic Association and its officers has been to rule out of order any resolutions taking positions on political issues;
> WHEREAS those rulings are based on a passage of Section III.3 of the Association's by-laws, which reads, "The Association as such will take no partisan attitude, nor will it commit its members to any position on practical economic matters;"
> WHEREAS this stricture, for the legal purpose of protecting the Association's tax-exempt status under Section 501 (c) 3 of the Internal Revenue Service code, need apply *only* to lobbying activities or political positions aimed directly at pending legislative matters;
> WHEREAS other professional academic associations regularly adopt positions on issues similar to those ruled out of order by this Association; and
> WHEREAS general political economic issues affecting the United States are too important for its leading economic association to preclude occasional comment:
> LET IT HEREBY BE RESOLVED:
> That the passage in Section III.3 of the Association's by-laws, which reads, "The Association as such will take no partisan attitude, nor will it commit its members to any position on practical economic matters" be deleted; and
> That it be replaced in Section III.3 of the by-laws by the following sentence: "The Association as such will take no position on any resolutions which pertain directly to pending legislation, or would otherwise jeopardize the Association's tax-exempt status."[1]

Gordon said that at a later point in the meeting he would introduce the following motion: The annual business meeting of the American Economic Association calls upon the Executive Committee to approve and set in motion the process necessary to affect the change in the language of the Certificate of Incorporation recommended in the resolution submitted by Robert Cherry and David Gordon. He expressed concern over the lack of attention paid to policy questions at the business meetings and called for a greater role for the Association in political affairs. Numerous political resolutions have been presented in the past and ruled out of order because of Section III.3 of the Certificate of Incorporation which proscribes the Association's taking partisan stands. He felt that serious discussion of political issues is needed, and that the Association should allow itself to take positions on economic questions.

Cherry spoke in favor of the resolution, saying that he was struck by the inflexibility of previous rulings by the chair on issue-oriented resolutions and that the present article is vague

[1]Although the resolution called for a change in Section III 3 of the Association's bylaws, the passage quoted and discussed is Section III 3 of the Association's Certificate of Incorporation. Much of the discussion from the floor involved the same confusion

Section III.3 of the bylaws reads. "The Association shall consist of the President, the President-elect, two Vice-Presidents, the Secretary, the Treasurer, the two Managing Editors, the two ex-Presidents who have last held office, and six elected members, provided the Secretary, the Treasurer, and the two Managing Editors shall not be entitled to vote in the Executive Committee's meetings."

and needs clarification. The elite of the profession has easy access to the media, but the elite does not necessarily speak for the profession. This biases the public's perception of the profession's opinions. He suggested that the Association could poll its membership on policy questions and simply announce the results of the poll.

George Tzanetakis opined that the Editor of the *American Economic Review* was violating the section of the Certificate of Incorporation in question because he encouraged and published articles on policy. Borts responded that the articles published do not constitute a statement by the Association They represent the opinions of the individual author.

After additional discussion, Gordon commented that several speakers seemed to think that the proposed change called for policy stands. Not so, he said. It simply allows the taking of positions. He moved the motion quoted above. Kraessel seconded it

Carolyn Shaw Bell offered the following friendly amendment: It is the sense of this meeting that the Executive Committee consider or cause to be reconsidered Section III.3 of the bylaws regarding no partisan attitude with a view toward clarifying the language and allowing the Association to take positions on questions of economic policy.[2] The mover and seconder of the original motion accepted this amendment.

Modigliani spoke against the motion, saying that we keep confusing free discussion of policies and its encouragement with the Association as such taking partisan positions on policy issues Passage of the proposed resolution would lead to less free discussion rather than more. Jim Robinson argued that members have numerous opportunities to discuss policy issues. The only possible objective to be served by passage of the resolution would be to trade on the Association's ability to influence public opinion and that should not be a goal of the Association. It may be appropriate for members as individuals, but not for the Association.

The motion was put to a vote and failed.

There being no further business, the meeting was adjourned at 11:45 p m.

C ELTON HINSHAW, *Secretary*

[2]See footnote 1 concerning the confusion between Section III 3 of the Certificate of Incorporation and Section III 3 of the bylaws

Minutes of the Executive Committee Meetings

Minutes of Meeting of the Executive Committee in Washington, D.C., March 19, 1976.

The first meeting of the 1976 Executive Committee of the American Economic Association was called to order in the Washington Hilton Hotel, Washington, D.C., at 9:10 a.m. on March 19, 1976. The following members were present: Franco Modigliani (presiding), Carolyn Shaw Bell, Barbara R. Bergmann, George H. Borts, Andrew F. Brimmer, Rendigs Fels, R. A Gordon, Walter W. Heller, C. Elton Hinshaw, Harry G. Johnson, Lawrence R. Klein, Mark Perlman, Edmund S. Phelps, Alice Rivlin, Paul M. Sweezy, and Burton A. Weisbrod. Donald F. Turner was present as the Association's Counsel. Present as members of the Nominating Committee were Kenneth Arrow (chair), William James Adams, Duran Bell, Marianne Ferber, Jack Hirshleifer, Leonard Rapping, and Vincent Tarascio. Present as guests for parts of the meeting were James Blackman, Gary Fromm, David Gordon, Irvin L. Grimes, Lloyd Reynolds, Fritz Machlup, and Diane Flaherty.

Minutes. The minutes of the meeting of December 27, 1975, were approved.

Report of the Secretary (Hinshaw). The Secretary reported that, pursuant to the guideline that normally firm site commitments should not extend beyond five years, he had informed the convention bureaus of the cities now holding space for the 1982 meetings that a decision will probably not be made until 1977. The cities being considered are New York, Montreal, Boston, and New Orleans. The Secretary reminded the Committee that once the place and dates of a particular annual meeting have been approved, they cannot be changed except in unusual circumstances, and a positive vote of two-thirds of the voting members of the Executive Committee would be required to effect a change. He reported that the total number of paid registrants at the 1975 Dallas meetings was 4,660. Of the people listed as program participants, approximately 420 did not preregister. Unless they registered after arriving at the meetings, the loss in revenue was $4,200. The Secretary and the Treasurer are in the process of renegotiating the Association's agreement with Richard D. Irwin, Inc., to distribute Volume XII (1970) of the *Index of Economic Articles* and are planning for the distribution of future volumes. The Committee discussed the issues involved in establishing the price of the *Indexes*.

Report of the Treasurer (Rendigs Fels). The Treasurer reported that financial statements for 1975 received from the auditors show an operating deficit of $56 thousand, a little less than the anticipated deficit of $62 thousand. The introduction of progressivity into the dues structure will probably result in a small surplus for 1976 rather than the expected deficit of $19 thousand in the budget adopted on December 27, 1975. The net worth of the Association at the end of 1975 was $92 thousand, little different from the figure previously reported. Since the net worth is hardly more than 10 percent of annual expenditures, the financial situation of the Association is precarious, though not desperate. The claim of the U.S. Post Office against the Association for $10 thousand has been settled for $3 thousand not counting lawyers' fees of $1,500. It was VOTED to authorize the Secretary, the Treasurer, and the Administrative Director to sign checks on behalf of the Association, and that two signatures be required for checks of $10 thousand or more. The authority previously granted to the Treasurer to borrow funds for the Association was not extended. It was VOTED to appropriate an additional $5,000 to support the activity of the Committee on the Status of Minority Groups in the Eco-

nomics Profession. It was understood that unexpended appropriations to the Committee would not automatically be carried forward, but that the Executive Committee would review "carry-over" requests each year It was requested that the Secretary write life members of the Association to solicit contributions.

Report of the Editor of the American Economic Review (Borts). The Managing Editor reported that he may have overestimated printing costs in the 1976 budget. He had projected a price increase for paper which has not and may not materialize

Report of the Editor of the Journal of Economic Literature (Perlman). The Managing Editor sought advice about whether the "Index of Authors of Selected Abstracts" section of the *Journal* should be reinstated He had originally estimated that the elimination of the "Index" would save $2,000 annually It now appears that the savings will be substantially less. It was VOTED to restore the "Index of Authors of Selected Abstracts" to the *Journal* The Managing Editor will handle the problem of making available the "Index" for the three issues in which it did not appear.

Report of the Director of Job Openings for Economists (Hinshaw). The Director reported that the number of subscribers had increased from the 1,660 previously reported to slightly over 1,800. The revenue projected for 1977 was based on the lower number of subscribers In the judgment of the Association's auditors, Arthur Andersen & Company, *JOE* is considered an unrelated trade or business, and any excess of revenue over expenses is subject to federal income tax.

Request for Special Appropriation (Machlup). Machlup reported on his pilot study of the extent to which journals are read He asked the Association to contribute toward the cost of postage for mailing the survey questionnaire to a sample of *AEA* members. It was VOTED to allocate up to $1,000 to help defray postage costs.

Committee on Foreign Honorary Members (Machlup). The Chairman reported that at present the Association's bylaws provide for the election of a maximum of 25 foreign honorary members. This number was adopted in 1946, at a time when the Association had 4,400 members. The membership now exceeds 18,500. The increase in membership and in the number of distinguished economists abroad warrants an increase in the number of foreigners to be elected honorary members of the Association. It was VOTED to submit to the members an amendment of the bylaws to remove the present restriction on the number of foreign honorary members and to authorize the Executive Committee to determine the number. It was understood that the Executive Committee would not increase the number of foreign honorary members unduly. It was VOTED that if the amended bylaw was approved by the members, the number of honorary members would be limited to 40 It was VOTED to elect the following as foreign honorary members: Herbert Giersch (Germany), Leif Johannsen (Norway), Janos Kornai (Hungary), Michio Morishima (U.K), Richard Stone *(U.K)*, Hirofumi Uzawa (Japan), and Herman Wold (Sweden) An additional slate of 10 was elected subject to the bylaws being amended.

Committee on U.S -Soviet Exchanges (Reynolds) The Chairman reported that the planned visit of *U S* economists to the *USSR* will probably occur June 7–21, 1976. The members of the *U.S* delegation and the subjects of their papers have tentatively been selected. The topic of the symposium will be "The Economics of Technological Progress "

Richard D Irwin, Inc. (Grimes). Grimes reported that he had been unsuccessful in his effort to remainder the inventory of volumes in the translation series. He offered to settle the Association's claim against Richard D. Irwin, Inc. for $17,000 and to dispose of the volumes according to the wishes of the Association. It was VOTED to establish a committee to investigate the settlement offer and to bring a recommendation to the next meeting of the Executive Committee

Federal Funding of Economic Research

(Blackman). Blackman reported on vicissitudes of federal funding of economic research. It was VOTED to establish an *ad hoc* committee to explore the problems and issues involved in the funding of research and to advise the Executive Committee whether the Association should take a more active role in seeking research support. The Committee is to report to the next meeting of the Executive Committee.

Nominating Committee (Arrow). Arrow reported the following nominees for offices in the 1976 election: Vice President (two to be chosen), Robert Eisner, Albert O. Hirschman, Leonid Hurwicz, and Anne O. Krueger; Executive Committee Members (two to be chosen), Marcus Alexis, Samuel Bowles, Robert Lampman, and Marc Nerlove. The Electoral College consisting of the Nominating Committee and the Executive Committee meeting together chose as nominee for President-elect Jacob Marschak, and as Distinguished Fellows Oskar Morgenstern and Herbert Simon.

Committee on Political Discrimination (Arrow). The Chairman reported that the Committee had received eight complaints against three institutions. Three of the complaints were filed jointly and one has subsequently been withdrawn, so that only five complaints were active. Two investigations have been completed and final reports transmitted to the individual faculty member, his or her academic department, the institution's president, and the American Association of University Professors *(AAUP)*. The Committee found no definitive evidence of political discrimination in either case. The Committee concluded that its original mandate will rarely, if ever, be feasible to execute. It was VOTED that the Committee should continue its present functions, that it should explore with the *AAUP* means of greater cooperation and joint procedures, and that a subcommittee be appointed to investigate whether the charge to the Committee should be broadened to include studies of the extent to which political discrimination occurs in hiring, promotion, salary, and research funding decisions.

Program Committee (Klein). The Chairman reported on his plans for the program of the 1976 meetings. In addition to sessions based on invited papers, there will be twelve sessions with contributed papers and special sessions on economic issues of the 1976 presidential campaign and the bicentennial of the *Wealth of Nations*.

Committee on Managing Editorships (Modigliani). The Chairman recommended that action on the Committee's report be postponed until the next meeting of the Executive Committee.

Date of Spring Meeting. It was agreed that the Executive Committee would meet on March 18–19, 1977, at a site to be determined.

The meeting was adjourned at 9:00 p.m.

Minutes of the Meeting of the Executive Committee in Atlantic City, New Jersey, September 15, 1976.

The second meeting of the 1976 Executive Committee was called to order at 10:20 a.m. on September 15, 1976 in the Haddon Hall Hotel, Atlantic City, New Jersey. The following members were present: Franco Modigliani (presiding), Carolyn Shaw Bell, Barbara R. Bergmann, George H. Borts, Rendigs Fels, C. Elton Hinshaw, Harry G. Johnson, Lawrence R. Klein, Mark Perlman, Edmund S. Phelps, Alice M. Rivlin, Paul M. Sweezy, and Burton A. Weisbrod Absent were Andrew F. Brimmer, Robert Aaron Gordon, and Walter W. Heller. Present as Counsel was Donald F. Turner. Present as guests for parts of the meeting were Marcus Alexis, David Gordon, Barbara MacPhee, and Lloyd G. Reynolds. Jacob Marschak was present for the entire meeting.

Minutes. The minutes of the meeting of March 19, 1976 were approved as circulated.

There was a correction of the minutes of the meeting of the Executive Committee in Washington D.C. on March 14, 1975. As recorded in the *American Economic Review, Papers and Proceedings*, May 1976, page 466, the change in Article V, Section 2 of the bylaws had already been approved by the members in the fall of 1975 and should have been omitted from the

list of amendments which the Executive Committee voted in December 1975 to submit to the members for approval this fall. The Secretary did not include it again among the revisions submitted to the members. The minutes of the March 14, 1975 meeting were approved as corrected.

Report of the Secretary (Hinshaw). The secretary reported that the 1977 annual meetings will be held at the New York Hilton Hotel in New York on December 28–30, with the employment service beginning operations one day earlier on December 27. The present schedule for subsequent meetings is: August 29–31, 1978 in Chicago; December 28–30, 1979 in Atlanta; September 5–7, 1980 in Denver; and December 28–30, 1981 in Washington, D.C.

Because this year's annual meeting comes at an early date in the academic year, arrangements have been made to have a supplementary job market at the O'Hare Hilton in Chicago during January 7–9, 1977.

As requested by the Executive Committee, the Secretary has written the life members of the Association congratulating them on their forecasting ability and asking them to contribute part of their savings on membership dues to the association. One positive response has been received.

The Secretary reported that the Association had entered into a new agreement with Richard D. Irwin, Inc. for the distribution of Volumes XI, XII, XIII, and XIV of the *Index of Economic Articles*. The agreement ends December 31, 1977.

Report of the Treasurer (Fels). The Treasurer presented preliminary financial reports for the first six months of 1976. (For details of the reports, see the Treasurer's Report in the *Proceedings*.) He projected a small surplus for 1976 as a whole, and stated that the Association's net worth is still uncomfortably low.

Report of the Editor of the American Economic Review (Borts). The Editor reported that beginning in 1977 the *Review* would be using photocomposition for typesetting. This may permit using the same size type for all articles and papers. As suggested by the Executive Committee at its December 27, 1975 meeting, he has commissioned an objective review article on the Means-Stigler controversy.

Report of the Editor of the Journal of Economic Literature (Perlman). On recommendation of the Managing Editor, the following persons were elected to the Board of Editors of the *Journal of Economic Literature:* Solomon Fabricant, William A. Miernyk, Michael J. Piore, and Barbara B. Reagan. It was VOTED to make the journal's data base available to students and faculty of the University of Pittsburgh as partial compensation for the services provided by that University to the Association. It was agreed that the Editor should explore means and determine the costs of making the data base available to others

Report of the Director of Job Openings for Economists (Hinshaw). The Director reported that expenses appear to be in line with full-year projections but that revenues will probably exceed the projections. *Job Openings for Economists* is now essentially self-supporting

U.S.-Soviet Exchanges (Reynolds) Reynolds reported on the June 1976 visit of ten *U.S.* economists to the *USSR* for a conference in Moscow on "The Economics of Technological Progress " It was concluded that the program has sufficient payoff to warrant its continuation, and that the Chairman should begin to seek a source of funding for another round of exchanges, i.e., a Soviet visit to the United States in 1977, followed by a *U.S* visit to the *USSR* in 1978. It was urged that the next *U.S.* delegation include minority group economists.

1977 Program (Marschak). The nominee for President-elect gave a brief review of plans for the 1977 program and reported that an organizer had already been approached for each session planned.

Committee on Economic Education. In the absence of G. L. Bach, Chairman of the Committee on Economic Education, Fels gave the Committee's report. The 1976 annual meeting marks the end of the Executive Committee's mandate of 1971 that there be at least one session on economic education at the annual meetings for the next five years, the papers to be

published on the same basis as other papers. It was VOTED to extend the mandate for another five years, papers to be published only if the Chairperson of the Committee on Economic Education judged them to merit publication.

Ad Hoc Committee on Federal Funding of Economic Research. In the absence of Stanley Lebergott, Chairman of the Committee, Modigliani summarized its written report. It was VOTED to appoint an *ad hoc* committee to evaluate the possible value of a standing committee on research funding and to seek to specify what the function and role of such a committee might be.

Representative to the Social Science Research Council (Klein). The Social Science Research Council *(SSRC)* has recently approved a change in the composition of its Board. As a result of the change, after a three year transition period the number of representatives from the Association will be reduced from three to one. It was agreed that the President write the *SSRC*, express qualms about the change in representation, and urge it to fill promptly the vacant staff position in economics.

Committee on Political Discrimination In the absence of F. Ray Marshall, Chairman of the Committee on Political Discrimination, Modigliani summarized a letter received from Marshall. David M. Gordon expressed disappointment over the slow response of this Committee to the Executive Committee's March 1976 request to investigate whether the charge to the Committee should be broadened to include studies of the extent to which political discrimination occurs in hiring, promoting, salary, and research funding decisions.

Committee on the Status of Minority Groups in the Economics Profession (Alexis). It was VOTED to carry forward any unexpended funds from this year's appropriation (Alexis estimated the expected carryover to be approximately $4,000) and to appropriate an additional $10,000 for the Committee with the hope and expectation that other sources of funds would be found and that this appropriation would not be used. It was agreed to consider cooperating with Educational Testing Service in obtaining data for purposes of evaluating graduate record examinations for minorities.

Committee on Managing Editors (Modigliani). The terms of both Borts and Perlman end in 1977. Borts has indicated that he is willing to serve only three more years; Perlman wants to serve two more years, but is willing to serve a third year if necessary. It was VOTED to extend the term of the Editor of the *American Economic Review* through 1980 and the term of the Editor of the *Journal of Economic Literature* through 1979.

International Economic Association Representative (Klein). It was VOTED to elect Modigliani to a five-year term as representative to the International Economic Association.

Richard D. Irwin, Inc. Settlement. In the absence of Brimmer, Modigliani reviewed the Committee's negotiations with Richard D. Irwin, Inc. concerning the Association's claim for $42 thousand in undistributed profits and accrued interest. It was concluded that negotiations should continue, that the accrued interest claim could be waived, and that the settlement of the claim should be for no less than the amount agreed to by the Executive Committee at this meeting.

Census Advisory Committee. In the absence of James R. Nelson, Chairman of the Census Advisory Committee, Bergmann reported the Bureau of the Census is under mandate to reduce the reporting burden by 5 percent, with the prospect of further reductions in the future. It appears that the Office of Management and Budget proposal for reducing the burden is simply to reduce the number of reports requested by the Bureau of the Census. Any trend in the direction of a weaker data base concerns the Association. It was VOTED to request members of the Advisory Committee to consult with appropriate officials of the Office of Management and Budget to determine the nature and extent of the "reporting burden" problem.

The meeting was adjourned at 5:30 p.m.

C. ELTON HINSHAW, *Secretary*

Interim Report of the Secretary For 1976

Annual Meetings. In 1977 the annual meetings will be held at the New York Hilton in New York City on December 28–30, with the employment service beginning operations on the 27th. The schedule for subsequent meetings is: August 29–31, 1978, in Chicago with headquarters at the Conrad Hilton Hotel; December 28–30, 1979, in Atlanta with headquarters at the Atlanta Hilton Hotel; September 5–7, 1980, in Denver with headquarters at the Denver Hilton Hotel; and December 28–30, 1981, in Washington, D.C. with headquarters at the Washington Hilton Hotel. No decision has been made for 1982. The Executive Committee has made a tentative decision to meet in San Francisco, December 28–30, 1983.

Employment Services. Because the 1976 annual meetings occurred at an early date in the academic year, it was decided to have a second placement service at the O'Hare Hilton in Chicago during January 7–9, 1977.

The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service under the direction of Mrs. Theresa Scholl. All economists looking for jobs and employers are urged to register. This is a placement service which maintains the anonymity of employers. The Association is indebted to Mrs. Scholl not only for the Registry but also for her and her staff's assistance and supervision of the employment service provided at the annual meetings.

Employers are reminded of their professional obligation to list their job openings in *Job Openings for Economists*.

Membership. The total number of members and subscribers, shown in Table 1, reached an all-time high of 27,753 as of June 30, 1976. At the end of 1975 the total number had regained the peak reached at the end of 1970 just before dues and subscriptions were doubled in price. It is too early to tell what impact the progressive dues structure adopted in 1975 will have on the numbers of members and subscribers.

TABLE 1—MEMBERS AND SUBSCRIBERS

	1975 (June 30)	1976 (June 30)
Class of Membership		
Annual	15,783	16,261
Junior	2,100	2,600
Life	369	400
Honorary	19	26
Family	325	336
Complimentary	443	447
Total Members	19,039	20,070
Subscribers	6,806	7,683
Total Members and Subscribers	25,845	27,753

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 104 as of June 30, 1976, compared to 204 for the full year of 1975. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

Visiting Economics Scholars Program. The Visiting Economics Scholars Program has continued under the direction of the Secretary. Its purpose is to facilitate visits by leading economists to smaller colleges emphasizing teaching. The colleges are expected to pay part or all of the costs of the visits; at a minimum they take care of the local expenses and travel of the visitor. During 1975–76 there were six such visits.

Committees and Representatives. Listed

below are those who served the Association during 1976 as members of committees or as representatives. Years in parentheses indicate the final year of the term to which they most recently have been appointed. On behalf of the Association, I wish to thank them all for their services.

Standing Committees

ADVISORY COMMITTEE ON STUDIES OF THE LABOR MARKET FOR ECONOMISTS

F. Ray Marshall, *Chairperson*
Barbara Reagan
T. Aldrich Finegan
Francis M. Boddy
John Siegfried (1978)
W. Lee Hansen (1978)
Rendigs Fels, *Ex Officio*

BUDGET COMMITTEE

Andrew F. Brimmer, *Chairperson* (1976)
Burton A. Weisbrod (1977)
Edmund S. Phelps (1978)
Franco Modigliani, *Ex Officio* (1976)
Lawrence R. Klein, *Ex Officio* (1977)
Rendigs Fels, *Ex Officio*

CENSUS ADVISORY COMMITTEE

James R. Nelson, *Chairperson* (1976)
Gardner Ackley (1976)
Armen Alchian (1976)
Andrew F. Brimmer (1976)
Anthony Downs (1976)
Jacob Mincer (1977)
George L. Perry (1977)
Lee Preston (1977)
Phyllis Wallace (1977)
Dale Jorgenson (1977)
Barbara Bergmann (1978)
Robert F. Lanzillotti (1978)
William Niskanen (1978)
Anne P. Carter (1978)
Richard Ruggles (1978)

COMMITTEE ON ECONOMIC EDUCATION

G. L. Bach, *Chairperson* (1976)
Henry H. Villard (1976)
Phillip Saunders (1977)
Elisabeth Allison (1978)

COMMITTEE ON HONORARY MEMBERS

Fritz Machlup, *Chairperson* (1976)
Lawrence R. Klein (1976)
Bent Hansen (1978)
W. Arthur Lewis (1978)
Leonid Hurwicz (1980)
Paul A. Samuelson (1980)

COMMITTEE ON HONORS AND AWARDS

Lloyd G. Reynolds, *Chairperson* (1976)
Gary Becker (1976)
Irma Adelman (1978)
Marcus Alexis (1978)
John Chipman (1980)
James W. McKie (1980)

COMMITTEE ON THE STATUS OF MINORITY GROUPS IN THE ECONOMICS PROFESSION

Marcus Alexis, *Chairperson* (1977)
George Borts (1976)
Andrew F. Brimmer (1976)
Alice Rivlin (1976)
James Tobin (1977)
Charles Z. Wilson (1977)

COMMITTEE ON POLITICAL DISCRIMINATION

F. Ray Marshall, *Chairperson* (1977)
William J. Baumol (1977)
John G. Gurley (1977)
Anne O. Krueger (1977)
Carl Stevens (1977)
Thomas E. Weisskopf (1977)
Gerald Somers (1978)

COMMITTEE ON PUBLICATIONS
Michael Lovell, *Chairperson* (1977)
Peter Diamond (1976)
Robert Lampman (1976)
John G. Gurley (1978)
Robert Ferber (1978)
Robert Gallman (1978)
C. Elton Hinshaw, *Ex Officio*

COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION
Barbara Reagan, *Chairperson* (1976)
Walter W. Heller (1976)
Isabel Sawhill (1977)
Janice Madden (1977)
Nancy H. Teeters (1977)
Margaret C. Sims (1978)
Franco Modigliani, *Ex Officio*

FINANCE COMMITTEE
Beryl W. Sprinkel, *Chairperson* (1976)
Robert Eisner (1977)
James Lorie (1978)
Rendigs Fels, *Ex Officio*

ECONOMICS INSTITUTE POLICY AND ADVISORY BOARD
Edwin S. Mills, *Chairperson* (1978)
Walter P. Falcon (1976)
Daniel Schydlowsky (1976)
Arnold Harberger (1977)
Richard H. Holton (1977)
Paul G. Clark (1978)
Carl Keith Eicher (1979)
Anne O. Krueger (1979)

JOINT COMMITTEE WITH THE ASSOCIATION OF AMERICAN LAW SCHOOLS
George J. Stigler, *Chairperson* (1976)
Alvin Klevorick (1976)

U.S.-SOVIET EXCHANGES
Lloyd G. Reynolds, *Chairperson* (1976)
Abram Bergson
John R. Meyer
Rendigs Fels, *Ex Officio*

Special Committees

AD HOC COMMITTEE ON FEDERAL FUNDING OF ECONOMIC RESEARCH (1976)
Stanley Lebergott, *Chairperson*
Milton Friedman
Gary Fromm
Zvi Griliches
Robert Solow

AD HOC COMMITTEE TO REVIEW NEW PROPOSED STANDARD OCCUPATIONAL CLASSIFICATION SYSTEM
H. Gregg Lewis, *Chairperson*
Victor Fuchs
Margaret S. Gordon
Michael Piori
Sherwin Rosen

COMMITTEE ON COMPUTERIZATION
John R. Meyer

COMMITTEE ON EDITORSHIPS (1976)
Franco Modigliani
R. A. Gordon
Lawrence R. Klein

NOMINATING COMMITTEE (1976)
Kenneth Arrow, *Chairperson*
William James Adams
Duran Bell
Marianne Ferber
Jack Hirshleifer
Leonard Rapping
Vincent Tarascio

NOMINATING COMMITTEE (1977)
Walter W. Heller, *Chairperson*
Nancy S. Barrett
Huey J. Battle
David M. Gordon
Bert G. Hickman
Irvin B. Kravis
Ralph W. Pfouts

COMMITTEE ON ELECTIONS (1976)
Ben Bolch, *Chairperson*
Barbara Haskew
C. Elton Hinshaw, *Ex Officio*

ATLANTIC CITY LOCAL ARRANGEMENTS COMMITTEE (1976)
Edward G. Boehne, *Chairperson*
Richard B. Benedict
Doris Burgess
John T. Callaghan
John J. Clark
Roger D. Collons
Thomas J. Doty
Marge Epps
Nancy Foltz
Robert C. Forrey
Vince Franco
Alan Gart
Alan Gersh
Shirley Goetz
Ben Han
A. Gilbert Heebner
Kathleen C. Holmes
Gerrold T. Jacobs
James Lawrence
Robert B. Lowry
Barbara MacPhee
Fred W. Malkin
Mary Murray
Louis Sauer
John B. J. Spraga
Joe Tumbler
Henry Watson
Bertham Zumeta

NEW YORK LOCAL ARRANGEMENTS COMMITTEE (1977)
Richard G. Davis, *Chairperson*
Dick Aspinwall
Peter Bakstansky
Lucian M. Caycy
Alice Christensen
Eleanor Johnson
Barbara MacPhee
Robert Schwartz
Violet Sikes
Thomas W. Synnott
Edward M. Syring, Jr.
Ingo Walter

Council and Other Representatives

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
Stephen Goldfeld (1976)

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SLAVIC STUDIES
Janet Chapman (1976)

AMERICAN COUNCIL OF LEARNED SOCIETIES
William Parker (1978)

AMERICAN POLITICAL SCIENCE ASSOCIATION—JOINT RESEARCH PROJECT ON CONFIDENTIALITY OF RESEARCH SOURCES
Gary Fromm

FEDERAL STATISTICS USERS CONFERENCE
John W. Kendrick (1977)

INTERNATIONAL ECONOMIC ASSOCIATION
Fritz Machlup (1976)
Abram Bergson (1978)

JOURNAL OF RESEARCH ON CONSUMER BEHAVIOR
Kelvin J. Lancaster (1978)

NATIONAL ARCHIVES ADVISORY COUNCIL—GENERAL SERVICES ADMINISTRATION
Robert Gallman (1978)

NATIONAL BUREAL OF ECONOMIC RESEARCH
Carl F. Christ (1978)

SOCIAL SCIENCE RESEARCH COUNCIL
Lawrence R. Klein (1976)
Guy Orcutt (1977)
Robert Eisner (1978)

Representatives of the Association on Various Occasions—1976

INAUGURATIONS

Dr. William A. Kinnison, Wittenberg University
Joseph T. Chao

Anthony J. Diekema, Calvin College
John O. Bornhofen

Aubrey Keith Lucas, University of Southern Mississippi
Lawrence B. Morse

Rev. Hugh F. Hines, Sienna College
Edwin J. Holstein

Joseph John Sisco, American University
Ransford W. Palmer

Robert S. Capin, Wilkes College
B. C. Dill

C. ELTON HINSHAW, *Secretary*

Report of the Treasurer For the Six Months Ending June 30, 1976

During the period 1969–75, the American Economic Association incurred deficits almost every year. The cumulated deficits, which totalled nearly half a million dollars, reduced the net worth of the Association to $111 thousand at the end of 1975, less than one-seventh of annual expenditures. The Executive Committee has taken a number of actions designed to eliminate the deficits, including the amendment to the bylaws approved by the membership last fall raising the dues of high-income members. This change went into effect on January 1.

A full year's experience with the new dues structure will be required to assess its effects

TABLE 1—AMERICAN ECONOMIC ASSOCIATION, REVENUES AND EXPENSES ACCRUAL BASIS, 1975–76
(Thousands of dollars)

	1976 Budget (12 months)	1975 Actual (12 months)	Actual, 6 Months Ending June 30, 1975	Actual, 6 Months Ending June 30, 1976
REVENUES				
Operating Income				
Membership dues	*	380	176	212
Subscriptions	*	237	118	122
Subtotal—dues and subscriptions	662	618	294	334
Job Openings for Economists	20	17	8	11
Advertising	69	64	32	36
Sales—back issues, etc	20	26	16	15
Sales—mailing list	32	33	17	17
Annual meeting	15	7	—	—
Other income	23	23	12	18
Subtotal—operating income	841	787	378	431
Investment Income				
Interest and dividends	40	37	17	15
Real capital gains (losses)	(60)	(85)	(32)	(17)
Subtotal—investment income	(20)	(48)	(14)	(2)
TOTAL REVENUE	821	739	364	429
EXPENSES				
Publications				
American Economic Review	214	196	96	96
Journal of Economic Literature and *Index*	277	253	120	138
Papers and Proceedings	56	51	49	58
Directory	50	61	30	25
Job Openings for Economists	15	22	7	6
Subtotal—publication expense	612	583	302	324
Operating and Administrative				
Salaries	120	104	46	56
Rent	8	8	4	4
Committees	25	23	13	7
Other	75	78	26	40
Subtotal—administrative expense	228	212	89	107
TOTAL EXPENSES	840	795	391	431
SURPLUS (DEFICIT)	(19)	(56)	(27)	(2)

*Not applicable

with confidence, but the results so far are encouraging. Table 1 shows the revenues and expenses of the Association for the first six months of 1976 compared to (1) the first six months of 1975, (2) the full year 1975, and (3) the budget for 1976. Revenues in the first half of 1976 were $65 thousand higher than in the corresponding period of 1975, an increase of 18 percent. More than half the increase came from dues, the consequence of the change in the dues structure. Expenditures were up $40 thousand (10 percent), of which $18 thousand was for the *Journal of Economic Literature* and the *Index*, $10 thousand for administrative salaries, and $9 thousand for the *Papers and Proceedings*. For the six months as a whole, the budget had a small deficit.

The results for 1976 as a whole are likely to be a little better than for the first half. For technical reasons, receipts from dues for the full year are regularly more than double the receipts for the first half. All or nearly all the expenditures for the *Papers and Proceedings* are incurred during the first half of the year. On the other hand, less than half the amounts budgeted for the *American Economic Review* and for committees was spent during the first half of the year. The Association is pressing a claim for $42 thousand against Richard D. Irwin, Inc. If any substantial part of this amount is collected during 1976, there will be a surplus for the year as a whole.

Table 1 is on an accrual basis. Table 2 shows recent developments in the cash position compared with the cash budget for the entire year 1976. For a number of reasons, there are large discrepancies between the two columns of the table. None of the printing bills for the four volumes of the *Index* scheduled for publication this year had come in by June 30 (line 4), so that the cash outlay for this item will be concentrated in the second half of the year. The very large discrepancy on line 5 is seasonal in nature. The Association regularly receives a large volume of payments of dues and subscriptions in December that are applicable to the ensuing year. At year end, they swell the amount of cash on hand without affecting the accrual budget and are recorded as "deferred income." During the year they are considered receipts from dues and subscriptions, quarter by quarter, so that the amount of deferred income declines heavily before recovering at the year end. The amount of receivables (line 6) has risen sharply in the last year, increasing $29 thousand in the first six months of 1976 to an aggregate of $82 thousand. Of this total, $60 thousand represents advances to the Economics Institute at the University of Colorado and to the Summer Program for Minority Students at Northwestern University, both of which are

TABLE 2—AMERICAN ECONOMIC ASSOCIATION SOURCES AND USES OF CASH, 1976
(Thousands of dollars)

	1976 Budget (12 months)	1976 Actual (6 months)
SOURCES (USES) OF CASH		
1. Budget surplus (deficit)	(19)	(2)
2 Addition to reserve for *Directory*	50	25
3 Real capital losses (gains)	60	17
4 (Manufacturing cost of *Index*)	(60)	0
5. Increase (decrease) in deferred income	39	(58)
6 Decrease (increase) in receivables	2	(29)
7 Increase (decrease) in accounts payable and accrued liabilities	12	34
8 (Furniture and equipment to be purchased)	(1)	-0-
9 Depreciation	*	1
10 Decrease (increase) in prepaid expenses	-0-	(9)
11 Restricted funds cash receipts less disbursements	-0-	(8)
INCREASE (DECREASE) IN CASH BEFORE TRANSFERS TO INVESTMENT ACCOUNT	83	(29)
Transfers to investment account	**	(65)
TOTAL INCREASE (DECREASE) IN CASH	**	(94)

*Less than 0.5
**Not applicable

sponsored by the American Economic Association. The amount of receivables can be expected to decrease during the second half of the year. The rise in accounts payable (line 7) is also a seasonal item, since the printing bill for the May issue was not paid until after June 30.

The cash position of the Association was strong enough to warrant the transfer of $65 thousand to the investment account during the first half of the year. Cash flow during the second half will be adequate to meet the needs of the Association.

RENDIGS FELS, *Treasurer*

Report of the Managing Editor American Economic Review

Because of the early appearance of the *Papers and Proceedings* issue, this report was completed prior to the end of the calendar year. Thus, it has not been possible to provide complete financial and publication data. In my report for 1977 to be given at the December meetings I shall provide a complete accounting for the years 1976 and 1977.

During the first 9 months of 1976, the number of papers submitted fell below the same period for 1975, 507 against 562. I estimate that 700 papers will be submitted for the whole of 1976. Table 1 shows the number of papers submitted and published for the past 20 years. There has been a gradual decline in submissions since 1970.

Table 2 shows the subject matter distribution of submitted and published papers for 1975 and 1976. The submissions cover comparable nine-month periods. The most popular fields are microeconomics, labor, monetary theory,

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1955–75

Year	Submitted	Published	Ratio of Published to Submitted
1956	242	48	20
1957	215	40	19
1958	242	46	19
1959	279	48	.17
1960	276	46	17
1961	305	47	15
1962	273	46	17
1963	329	46	.14
1964	431	67	.16
1965	420	59	.14
1966	451	62	.14
1967	534	94	18
1968	637	93	15
1969	758	121	16
1970	879	120	14
1971	813	115	14
1972	714	143	20
1973	758	111	15
1974	723	125	17
1975	742	112	15
1976	700 (est.)	113	16 (est.)

TABLE 2—SUBJECT MATTER DISTRIBUTION OF SUBMITTED AND PUBLISHED MANUSCRIPTS, 1976 AND 1975

	Submitted		Published	
	1976	1975	1976	1975
	(January 1—September 30)			
General Economics and General Equilibrium Theory	14	8	2	1
Microeconomic Theory	79	67	21	18
Macroeconomic Theory	45	55	11	9
Welfare Theory	42	63	7	17
Economic History, History of Thought, Methodology	5	6	—	—
Economic Systems	27	18	—	3
Economic Growth, Development, Planning Fluctuations	19	20	10	3
Economic Statistics	18	20	2	6
Monetary and Financial Theory and Institutions	47	49	18	8
Fiscal Policy and Public Finance	28	21	3	6
International Economics	41	58	14	11
Administration, Business Finance	18	14	2	4
Industrial Organization	28	24	4	4
Agriculture, Natural Resources	10	7	3	2
Manpower, Labor, Population	66	88	10	18
Welfare Programs, Consumer Economics, Urban and Regional Economics	20	44	6	2
Total	507	562	113	112

macroeconomics, welfare theory, and international economics. The distribution remains remarkably stable from year to year, as an examination of past reports will show.

It is too soon to know the reasons for the decline in submissions. One possibility is the introduction of page charges. If this is the case, it may be due to a misapprehension on the part of authors. The charge of $25 per journal page is payable by the institution or granting agency supporting the research. Payment of the charge is not a prerequisite for publication, nor are authors expected to pay the charges themselves. Moreover articles are accepted without prior knowledge of financial support the author may have. If the author or his/her institution cannot pay the charge, it is waived. Beginning in March 1977, this statement of policy on page charges will appear in every issue of the *Review* so that any possibility of misunderstanding should be eliminated.

Another possible reason for the decline in submissions is the delay inherent in the screening and refereeing process. Authors have a wider range of journals to choose for submissions and may find that delays are smaller with some of the newer journals. The two-step review process is designed to cut down the number of papers that are rejected solely at the discretion of the managing editor. Each paper is read first by a screener to determine its potential suitability for inclusion in the *Review*. Screeners are paid $10 per manuscript to read the paper, determine if it is a serious piece of scholarly work appropriate for the *Review*, write a brief summary, and suggest potential referees. Approximately 40 percent of the papers are rejected after they are screened and read by the managing editor. The remainder are refereed, and of those approximately 25 percent are published. Many rejected papers are reworked and submitted again either to the *Review* or other journals.

There have been delays in the refereeing of papers, and some referees have suggested that they be paid. It is believed that a small fee (say $25) would serve as a reward to speed up the return of referee reports. Payment for refereeing has begun at some other journals, and it may be possible to introduce a similar practice at the *Review*. However payment for refereeing would require an additional submission fee, and this matter is under examination. The present submission fee of $15 for members ($30 for nonmembers) serves to finance the costs of screening, but it would have to be increased to cover fees as well.

A second aspect of the refereeing process is under examination. It is the role of the Board of Editors of the *Review*. At the moment the Board consists of eighteen members. Their names are printed on the contents page of every issue. Members of the Board are chosen by the managing editor with the approval of the Executive Committee of the Association. The Board functions mainly as a group of super referees. Each Board member agrees to read and referee up to eight papers a year in his/her field of specialization. The Board and the editor are chiefly responsible for maintaining the quality of the *Review*. However, Board members do not make final editorial decisions, and their policy role is advisory. From one point of view this limited role of the Board might be considered an underutilization of talent; from another viewpoint it is a way of getting very valuable services in small quantities from some very able and busy people. I am presently discussing with the Board possible methods of using their services and abilities in a wider range of editorial activities.

Invited Papers

I have continued the practice of inviting papers on policy issues. The purpose is to stimulate discussion and to erode the apparent unwillingness of most economists to do policy research. This year the September issue of the *Review* contained four papers on British inflation. The authors are R. J. Ball and T. Burns, David Laidler, Marcus Miller, and John Williamson and Geoffrey Wood. I expect these papers will generate a good deal of comment by British and American economists, and space will be devoted to such comments in later issues. Next year the *Review* will publish three papers on the U.S. Social Security System, by Michael Boskin, Martin Feldstein, and Paul Samuelson.

I am still waiting for comments on the Radical Critique of the Council of Economic Advisors Report, an invited paper which appeared De-

TABLE 3—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING OF THE FIRST THREE ISSUES OF 1976

Copies	Copies Printed	Pages		Cost		
		Net	Gross	Issue[b]	Reprints[c]	Total
March	29,200	252	304	$34,767 15	$619.01	$34,148.14
June	29,000	214	236	29,976.26	332.05	29,644.21
September	27,500	264	296	37,465 00[a]	330.00[a]	37,135.00[a]
Three-Issue Total	85,700	730	836	$102,208.41[a]	$1,281.06[a]	$100,927.35[a]

[a]Estimate.
[b]Includes allocated cost of preparing mailing list.
[c]Credit resulting from charges to authors for additional reprints

cember 1975. Despite a good deal of verbal comment which I was privileged to receive, this paper has not generated any significant attention from authors.

Expenses—Printing and Mailing

The 1976 printing and mailing costs will remain within the limits set by our budget. Increased costs over 1975 will occur because of higher paper prices and higher mailing charges, but they have been anticipated. I expect however that 1977 will see a large jump in printing charges. We will be subjected to higher composition and manufacturing costs in our next contract with the printer. As soon as these are known with some accuracy we shall compare them with the competition.

Board of Editors

Six members of the Board of Editors will complete their terms at the end of this year: Martin Feldstein, Bent Hansen, Jerome Stein, S. C. Tsiang, Finis Welch, and Marina v.N. Whitman. I wish to thank them for their high professional standards, work, and cooperation.

I wish to express my thanks to the continuing members of the Board of Editors: Irma Adelman, David Baron, Robert Barro, Laurits Christensen, Eugene Fama, Robert Gordon, David Laidler, James Melvin, William Nordhaus, Stephen Resnick, Anna Schwartz, and Frank Stafford.

I shall submit the names of six new members of the Board to the Executive Committee at its meeting in March 1977.

Acknowledgments

I should like to thank my associates for their cooperation and patience: Wilma St. John for her fine work as assistant editor; Carol Chapin our editorial assistant, and Jill St. John our secretary.

The following assisted me this year as editorial consultants: Michael Abbott, Ann Bartel, Farrell Bloch, George Borjas, Linda Edwards, Ronald Ehrenberg, Dennis Epple, Marie-Therese Flaherty, Donald Frey, Carl Gambs, Frank Gollop, James Hanson, Edward Lazear, Charles Lieberman, David McNichol, Ronald Oaxaca, Robert Rohr, Mark Rosenzweig, Thomas Russell, Jose Alexandre Scheinkman, Steven Shavell, John Shoven, Robert Smith, and Hal Varian.

In addition to the members of the Board and the editorial consultants, I have sought and received the assistance of a large number of economists during the course of the year. I wish to thank them for their cooperation and high standards in reading and evaluating manuscripts. They have eased the work-load that would otherwise fall on the Board of Editors. The following have assisted as referees.

H. Aaron
B. Aghevli
N. Aitken
P. Allen
J. Anderson
O. Ashenfelter
L. Auernheimer
E. Bailey
M. Bailey
R. Baldwin
E. Baltensperger
N. Barrett
J. Barron
M. Barth
Y. Barzel
R. N. Batra
W. Baumol
J. Behrman
L. Benham
G. Benston

U. Ben-Zion
T. Bergstrom
E. Berndt
S. Berry
T. Bertrand
H. Binswanger
F. Black
S. Black
R. Blackhurst
M. K. Block
A. Blinder
Z. Bodie
T. Borcherding
K. Boulding
R. Boyer
W. Branson
F. Brechling
A. Brillembourg
W. Brock
J. Buchanan
G. Butters
P. Cagan
G. Calvo
T. Cargill
R. Carson
F. Casas
C. Cathcart
R. Caves
W. Chang
P. Cheng
J. Chipman
G. Chow
C. Christ
P. Clark
C. Clotfelter
P. Coelho
R. Coen
W. Comanor
B. Conley
P. Cooper
J. Cox
R. Craine
M. Crew
M. L. Cropper
K. Davis
E. Davis
O. Davis
R. H. Day
R. Deacon
M. DeGroot
G. DeMenil
E. Denison
A. Denzau
P. Desai
J. DeSalvo
A. DeVaney
P. Diamond
B. Dieffenbach
W. E. Diewert
A. Dixit
P. Doeringer
F. T. Dolbear
M. Dooley
R. Dorfman
R. Dornbusch
G. Douglas
S. Dresch
G. Eads
C. Eaton
R. Eckaus
R. Ehrenberg
I. Ehrlich
R. Eisner
B. Ellickson
J. W. Elliott
E. Elton
D. Epple
T. Epps
R. Falvey
L. S. Fan
G. Faulhaber
A. Feldman
W. Fellner
T. A. Finegan
J. M. Finger
S. Fischer
P. Fishburn
A. Fisher
J. Flanders
A. M. Freeman
R. Freeman
A. Friedlaender
J. Friedman
I. Friend
E. Furubotn
R. Gallman
J. Geweke
J. F. Giertz
A. Gifford
L. Girton
C. Goetz
R. Goldfarb
S. Goldfeld
M. Gordon
R. A. Gordon
J. Gould
H. Grabowski
H. Gram
M. Greenhut
R. Grieson
H. Grossman
M. A. Grove
T. Groves
H. Grubel
M. Guitian
J. Gwartney
J. Hadar
R. Hall
K. Hamada
M. Hamburger
B. Hamilton
W. L. Hansen
J. Hanson
D. Harris
M. Harris
R. Hartman
J. Hartwick
R. Haveman
T. Havrilesky
J. M. Heineke
W. Heller
D. W. Henderson
J. V. Henderson
J. Hirshleifer
O. Hochman
W. Holahan
C. Holt
D. Holthausen
H. Hori
A. Horowitz
T. Horst
P. Howitt
E. Howle
E. P. Howrey
C. Hulten
P. Isard
D. Jaffee
E. James
M. Jensen
G. Johnson
C. R. Jones
E. B. Jones
R. Jones
R. W. Jones
P. Joskow
M. Kamien
E. Karni
P. Kenen
J. Kennan
M. Khan
R. Kihlstrom
B. Klein
J. Kmenta
L. Kochin
T. Koizumi
R. Komiya
G. Kopits
M. Kosters
M. Kreinin
A. Krueger
D. Laidler
J. Laitner
R. Lampman
W. Landes
H. Lapan
L. Lau
L. Lave
E. Leamer
J. Ledyard
A. Leibowitz
A. Leijonhufvud
H. Leland
S. LeRoy
H. Levy
K. Lewis
C. Lieberman
H. Liebhafsky
C. Link
R. Lipsey
D. Logue
J. Lothian
B. T. McCallum
D. McFadden
R. Mackay
S. Maital
J. Makin

M. Manove
E. Mansfield
J. Marchand
J. Marshall
S. Masters
F. Mathewson
P. Mattila
W. Mayer
D. Mayers
A. Meltzer
R. Merton
R. Meyer
R. Michael
P. Mieszkowski
L. Mirman
F. Mishkin
H. Mohring
M. Montias
J. Morgan
M. Morishima
S. Morley
J. Moroney
D. Mueller
G. Mumy
Y. Mundlak
M. Mussa
R. Muth
K. Nagatani
P. Neher
C. Nelson
M. Nerlove
D. Newbery
J. Newhouse
P. Newman
Y. K. Ng
A. Nichols
J. Niehans
Y. Niho
R. Noll
W. Oakland
W. Oates
J. Ohls
B. Okner
E. Olsen
M. Olson
J. Ordover
L. Orr
K. D. Osborne
J. Ostroy
H. Pack
M. Paglin
M. Parkin
R. Parks
D. Parsons
P. Passell
M. Pauly
S Peltzman
S. Perrakis
H. C Petersen
F. Peterson
E Phelps
L. Phillips
R. Pindyck
D. Pines
M Piore
C. Plott
M. Polinsky
R Pollak
H Pollakowski
W. Poole
E. Prescott
J. Quigley
T. Rader
L Rapping
R. Rasche
E. Ratledge
A. Razin
E Rodriquez
R. Rohr
H. Rose
S. Rose
S. Rose-Ackerman
S. Rosefielde
M Rothschild
M. Rubinstein
L. Ruff
R. Ruffin
K. Russell
W. Russell
J. Rutledge
H. Ryder
S. Salant
A. Sandmo
A. Santomero
T. Sargent
M. Sarnat
K. Sato
R. Sato
I. Sawhill
R Schmalensee
R. Schuler
P. Schultz
C. Schultze
W. Schulze
M. Schupack
A. Schweinberger
J. Seater
A. Sen
E. Seskin
P. Shapiro
W. Sharpe
R. Shearer
E. Sheshinski
R. Shishko
F. Shupp
J Siegel
W. Silber
D. Sjoquist
E. Silberberg
C. Sims
K. Smith
L. Smith
V. Smith
R. Solow
R. Spann
M. Spence
W. Springer
R. Starr
D. Starret
L. Steinhaver
P. Stephan
D. Stewart
G. Stigler
J. Stiglitz
B. Stigum
M. Strober
W. Stubblebine
W. A. Stull
A. Swoboda
E. Tanner
P. Taubman
J. Taylor
L. Telser
M. Teubal
L. Thurow
N. Tideman
P. A. Tinsley
J. Tobin
R. Toikka
G. Tolley
R. Tollinson
E. Tower
E. Truman
S. Turnovsky
D. Usher
H. Varian
E. C. H. Veendorp
J. Vernon
W. Vickrey
D. Vining
M. Visscher
H. Votey
P. Wachter
N. Wallace
R Waud
L. Waverman
W Weber
B. Weisbrod
R Weiss
M. Weitzman
S. Wellisz
W Wheaton
L White
R. Wichers
J. Williamson
P. Williamson
R. Willig
D. Wise
J. Witte
R. Wong
Y. H. Yeh
J Yellen
E. Zabel
E. Zajac
R. Zeckhauser

GEORGE H. BORTS, *Managing Editor*

Report of the Managing Editor Journal of Economic Literature

This report, prepared for the September 1976 meetings at Atlantic City, does not contain the usual financial information; it is not feasible to make projections of expenditures so far in advance of the end of the budget year.

Table 1 illustrates the projected allocation of space in the *Journal of Economic Literature (JEL)* for 1976 as well as the comparisons for the years 1972 through 1975 Table 2 classifies the material by subject both for the 1976 issues and the totals for the period 1969 through 1976. And, finally, Table 3 classifies the material by technical difficulty. Again, a change has been made to indicate the direction during the past year (that is, 1976).

Members will note that we published three, not four, survey articles during 1976, and that we published eight, not four, essays on the literature in subfields. Normally we publish one article analyzing bibliographic development, as such; this year none was published.

We have, in various stages of completion, commissioned survey articles on the microfoundations of macroeconomics, the discussion of an impending American capital shortage, the welfare implications of national income accounting, the literature on the role of population in economic development, the literature on testing of economic hypotheses, possibly one on the economics of forecasting. Apparently there is considerable interest in the profession on expanding some of our survey articles into book length monographs. Incidentally, if any member has a particular interest in bringing to my attention (and through me, of course, to the attention of the Board of Editors) the need for a survey article in any (heretofore neglected) subfield, I would be grateful if this information were communicated to me by letter

During 1976 we have managed to bring to

TABLE 1—QUANTITATIVE ANALYSIS OF CONTENTS, *JEL*, 1972 THROUGH 1976
(Number of pages in parentheses)

	1972		1973		1974		1975		1976*	
	No	Pages	No	Pages	No	Pages	No	Pages	No	Pages
Survey articles	3	(78)	4	(144)	3	(102)	3	(119)	3	(105)
Essays on subfields	5	(129)	2	(28)	5	(99)	5	(100)	8	(185)
Review articles	5	(39)	—	—	1	(5)	—	—	—	—
Articles about economic literature	1	(19)	2	(38)	—	—	1	(11)	—	—
Communications	4	(13)	10	(26)	13	(71)	12	(36)	3	(10)
Books annotated	1,209	(257)	1,214	(239)	1,211	(229)	1,203	(223)	1,200	(260)
Books reviewed	160	(241)	175	(259)	168	(239)	183	(282)	185	(280)
Journal issues listed and indexed	849	(140)	1,011	(185)	986	(180)	908	(177)	900	(175)
Number of individual articles	5,387	—	7,218	—	7,360	—	6,788	—	6,500	—
Subject index of *Journal* articles		(263)		(357)		(338)		(349)		(315)
Abstracts of articles	1,415	(313)	1,906	(407)	1,645	(312)	1,637	(331)	1,500	(310)
Total pages**		(1,572)		(1,748)		(1,671)		(1,700)		(1,640)

*Preliminary; will be revised prior to 1977

**Includes, in addition to listed pages, classification systems, table of contents, indices, journal subscription information, etc

TABLE 2—CLASSIFICATION BY SUBJECT, 1969–76 (INCL.)

	1976		1969–76 All articles
	Commissioned Survey	Creative Curmudgeon Essays	Total*
01 General	—	—	6
02 Theory	1	2	20
03 Thought (Methodology)	—	2	21
04 Economic History	—	1	3
05 Comparative Systems	—	—	4
11–12 Growth & Development	—	—	6
13 Stabilization	—	1	1
21–22 Econometric, Statistical Theory, Statistics	—	—	3
31 Monetary Economics	—	1	4
32 Fiscal Economics	—	—	4
40–44 International Economics	—	—	11
50 Managerial Economics	—	—	1
60 Industrial Organization, Industrial Regulation		—	1
70 Agricultural and Resource Economics	1	—	2
80 Labor Economics	1	1	6
90 Applied Welfare Economics, Regional Economics	—	—	6
Totals	3	8	99

*Includes all review articles on books, general essays on all literature

TABLE 3—CLASSIFICATION BY TECHNICAL DIFFICULTY, 1969–76 (INCL.)

	1976		1969–76 Totals
	Surveys	Creative Curmudgeon Articles	Surveys Creative Curmudgeon Articles Others*
Most Difficult	—	1	19
Some Difficulty	3	4	45
Not Difficult	—	3	35
Totals	3	8	99

*Review articles or books and general essays on all literature, excludes very short communications

press *Annual Indexes* for the years 1970, 1971, 1969, and 1972. These volumes are processed only in part as a byproduct of the quarterly *Journal of Economic Literature*. The *JEL* articles are those appearing in the quarterly *JEL* arranged by year of publication in the original *Journal;* articles from books of collected essays and documents of the relevant year are added. Each volume contains many more entries than was the case in the previous series of the *Annual Index of Economic Articles*. We hope that it will be possible to produce during 1977 the *Annual Indexes* for 1973 and 1974. Thereafter, that is starting with 1978, one *Annual Index* will be produced each year. Members may ask why there is a two- to three-year delay in the publication of an annual index. The explanation is quite straightforward. It would not be unusual for us to receive the last December 1976 issue of a particular journal as late as November 1977. We cannot process any annual volume until the last issue of the journals covered has arrived and

been processed for the *JEL* quarterly. Thus, it is entirely possible that an issue of some journal *X* arriving in our office in November 1977 would not be processed for a *JEL* quarterly issue until May or June 1978. Putting together the *Annual Index* involves, as I note below, a great deal of checking. Six months are often spent on the process. Consequently, actual publication for the 1976 *Annual Index* might not occur until early 1979—*quod erat demonstrandum*. We have managed to put out these publications with no additions to our full-time staff.

Of pressing interest at the time of this report is my concern with the current selection of journal coverage for the indexing operation. We cover over 200 journals. Most of them were a legacy from an earlier incarnation, the *Journal of Economic Abstracts*. In the last five years a great many new journals have come into existence. I have been concerned that the tradeoff between completeness of coverage and economy of the operation be rational. The matter of whether to permit additional journals for listing (and to permit the even rarer cases of additional journals for abstracting) has to be approached several ways. In all likelihood I will probably be publishing an Editor's Note on some empirical findings which we are in the process of studying at this time. In any event, I hope that by the beginning of 1977 I will be more confident of the basis of my selection of journals for listing and abstracting.

The Chancellor and the Dean of the Faculty of Arts and Sciences of the University of Pittsburgh have again this year allocated some University of Pittsburgh support to the *Journal*. Their willingness to do so, particularly in this period of retrenchment and tight budgets, illustrates an understanding of and a devotion to scholarly work in the economics discipline. I have regularly thanked them privately and take this opportunity to do so again publicly.

Four members of the Board of Editors have completed their terms. I wish to convey to them publicly (although I have already done so privately) my great appreciation of their tremendous help. Marcus Alexis (Northwestern University), Anne P. Carter (Brandeis University), Tibor Scitovsky (Standford University), and William Vickrey (Columbia University) have been exactly what every managing editor wants. The other members of my Board have also been superb colleagues. I look forward to my continuing association with them. I have nominated several people to replace the departing four; it is, however, premature for me to announce the names in this report. (The procedure is for me to nominate a committee which meets after this report has been prepared to confirm my nominations.)

I wish also to thank the following economists (plus four who have chosen to remain anonymous) for advice and assistance in the commissioning, refereeing, and revising of articles:

Armen Alchian	Michael Rothschild
Orley Ashenfelter	George L. S. Shackle
Abram Bergson	Henry Wallich
Martin Bronfenbrenner	Burton Weisbrod
Ray Marshall	Basil Yamey
Allan Meltzer	

Finally, the *Journal* staff, in its offices in Pittsburgh, has done simply superb work during this year. We have managed to bring out our issues on schedule while at the same time preparing four *Indexes* for publication with two more in the early stages of the publication process. The Associate Editor, Mrs. Naomi Perlman, has done the lion's share in creating and implementing a detailed 4-digit subject index system, in developing a topical guide to the classification schedule, in working out many programming problems with the printer, and in overseeing what seems to be an almost endless checking process. The Assistant Editor, Mrs. Drucilla Ekwurzel, seems to me also to have achieved the near impossible in supervising the proofreading tasks. Of course, we have had considerable help from Mrs. Lyndis Rankin (the principal secretary), from Mrs. Margaret Yanchosek (who handles the record keeping involved in the indexing process, from Miss June Cox (who, after many years on the *Journal*, has taken another assignment at the University of Pittsburgh), and from Mrs. Carolyn Simon, who is Miss Cox's successor.

MARK PERLMAN, *Managing Editor*

Report of the Director Job Openings for Economists

In the first four 1976 issues of *Job Openings for Economists (JOE)*, employers listed a total of 949 vacancies—a 4 percent increase over the same period last year. Of these, 746 were jobs not previously listed. About 62 percent of the new positions were classified as academic; the remainder were nonacademic. The same four issues in 1975 carried a total of 913 vacancies of which 642 were new. About 66 percent of the new positions advertised in 1975 were classified as academic. Table 1 shows the total listings, total jobs, new listings, and new jobs for each of the first four 1976 issues.

As in 1975, universities with graduate programs and 4-year colleges were the major advertisers of jobs—48 and 31 percent of all employers. Consulting and research firms were the largest advertisers for nonacademic positions. Table 2 shows the number of employers by type for each issue.

TABLE 1—JOB LISTINGS FOR 1976

Issue	Total Listings	Total Jobs	New Listings	New Jobs
	Academic			
February	111	229	73	143
April	72	115	56	90
June	47	82	42	73
August	86	169	78	155
Subtotals	316	595	249	461
Issue	Nonacademic			
February	19	75	13	49
April	17	64	11	45
June	24	58	21	47
August	22	157	16	144
Subtotals	82	354	61	285
Totals Academic and nonacademic	398	949	310	746

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1976

Issue	Four-Year Colleges	Universities with Graduate Programs	Junior Colleges	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	44	66	1	4	—	6	4	5	—	130
April	35	36	1	2	1	3	4	7	—	89
June	18	28	1	7	1	3	1	6	6	71
August	25	61	—	4	2	5	2	7	2	108
Totals	122	191	3	17	4	17	11	25	8	398

General economic theory continues to be the field of specialization most in demand. Table 3 lists the number of citations by 1976 issue for each of the broader field classifications. General economic theory led with 21 percent of the total number of citations, followed by business administration, finance, marketing and accounting (13 percent), monetary and fiscal (12 percent), welfare and urban (11 percent), and econometrics and statistics (10 percent). These same specialities were cited in the same order during 1975, and they each had approximately the same percentage share of total citations as last year.

At the beginning of this year, the projected deficit for *JOE* was $500. Revenues were estimated to be $19,350, and expenses to be

TABLE 3—FIELDS OF SPECIALIZATION CITED

Field[a]	February 1976	April 1976	June 1976	August 1976	Totals 1976
General economic theory 000	75	49	39	73	236
Growth and development 100	24	13	8	22	67
Econometrics and statistics 200	43	25	22	29	119
Monetary and fiscal 300	50	23	17	44	134
International economics 400	21	13	13	18	65
Business administration, finance, marketing and accounting 500	57	35	21	32	145
Industrial organization 600	22	15	24	27	88
Agriculture and natural resources 700	12	9	11	16	48
Labor 800	23	18	17	22	80
Welfare and urban 900	42	20	14	45	121
Related disciplines A00	9	5	0	0	14
Administrative positions B00	9	8	2	8	27
Totals	387	233	188	336	1,144

[a]Fields of specialization codes are from the *Journal of Economic Literature*.

$19,850. The subscription revenue projection was based on 1,662 subscribers and was conservative. *JOE* now has slightly over 2,000 subscribers. Expenses appear to be in line with the projection. An increase in subscription rates or the initiation of a charge for listing does not appear necessary at this time to keep *JOE* self-supporting.

C. ELTON HINSHAW, *Director*

Report of the Committee on the Status of Women in the Economics Profession

It comes as no surprise to economists that sex discrimination depresses women's wages or salaries, and the concomitant underemployment of women lowers total economic production. Great gains in productivity could be realized if barriers to occupational segregation by sex were removed.

The surprise to some economists has been the pervasive and subtle nature of sex discrimination within our own profession, and the long-run nature of any solution. The surprise to women economists has been the extent to which advances toward equality of opportunity create resistance towards further movement. Within the internal labor markets of business, government agencies, and universities, it can be seen that many men these days are apprehensive and perceive reverse discrimination when in fact women are making little or no progress. In this setting, the stand of our professional association to obviate sex discrimination and this Committee's efforts to help the process remain a major priority.

Economics is a field that has typically been male-dominated. We economists in colleges and universities have urged all undergraduate students, female as well as male, to take principles of economics, problems courses, and even more advanced work to enhance their general education. As public interest in economic policy has grown, the number of students in our classes has increased. We have stressed the value of studying economics as a foundation or complement for further specialization in areas such as business, law, and engineering as well as for research and teaching of economics. We have sought female majors as well as male. The proportion of women among undergraduate majors in economics has grown somewhat in recent years, but is still low. In 1974–75, of the B.A. degrees in economics, 22 percent were earned by women compared with 16 percent in the two previous years. The proportion of Ph.D. degrees in economics earned by women was only 11 percent in 1974–75, which represents only a slight increase in recent years (see Reagan, 1976). Even though we want to end occupational segregation by sex, it is not enough for educators to advise young women to go into fields of specialization that are atypical for women such as economics. Increasingly it is clear that support systems for the young women in atypical fields must be developed, and barriers to their career growth must be removed. Our Committee has worked on various aspects of these issues for the economics profession.

It has been now five years since the American Economic Association *(AEA)* created a Committee on the Status of Women in the Economics Profession *(CSWEP)*. This has been a hard-working group of women and men dedicated to carrying out the Association's mandate to a) support and facilitate equality of opportunity for women economists in all aspects of economists' professional activities, and b) help eradicate any institutional or personal discrimination against women economists. A brief summary of the activities for the five-year period with notes on the increments made this year by the Committee is given below.

The Committee membership has rotated to include representatives from various segments of the country, from business, universities, government, and private research organizations; and within universities from economists currently working in departments of economics, agricultural economics, and schools or departments of business, and at various ranks from as-

sistant professors to full professors and administrators.[1] I want to thank these members for their commitment and service to the Association on this committee.

The Resolutions under which the Committee has been operating were adopted at the December 1971 meetings of the *AEA* in New Orleans and published in the Appendix to the Minutes of the Annual Meeting, pages 473–74 of the May 1972 *American Economic Review*. For the first three years of the Committee's life, the Association funded its activities with two grants from the Ford Foundation to enable the Committee to undertake at a rapid rate the various projects in line with its charge from the *AEA*. It was understood in accepting the start-up grant that if the experiment worked, the *AEA* would continue the activities and take over their financing. *CSWEP* has been a Standing Committee of the Association for nearly three years now, and basic operations have been fully financed by the Association for the last two years. In addition, grants were received from the Carnegie Corporation of New York and from the German Marshall Fund of the United States for special activities of the Committee.

The major guideline developed by *CSWEP* for its activities and follow-up to these activities by Association members is a call for good-faith efforts to redress the present low representation of women in the economics profession. To effect this, following the December 1971 mandate of the Association, *CSWEP* has developed the following areas for its activities. We have identified women economists throughout the country, encouraged membership in the *AEA*, and increased the effective supply of women economists. We have tried to help meet the demand for women economists for job openings and professional activities such as committee work and program participants, and have tried overtly to stimulate such demand. We have worked to improve the workings of the labor market for economists. We have built an informal network among women economists across the country and provided informal support for our women colleagues who have often felt isolated in our profession. This is a necessary step to increase the supply of women economists. A steady, unsolicited stream of letters over the past five years attests to the appreciation of our women colleagues for *AEA*'s efforts through *CSWEP* to provide this informal support. We have counseled on request with individual women and their male colleagues (particularly department chairmen). Guidelines for implementing the affirmative action resolutions passed by *AEA* in December 1971 were developed by *CSWEP*, subjected to public debate, and published (see Boulding and Reagan). We have worked to avoid duplication and to help centralize the efforts of all women's organizations by having a close liaison with the Federation of Organizations for Professional Women. Perhaps most important of all, *CSWEP* has conducted and published ongoing research on the status of women economists. We have encouraged research on the broader issues of the role of women in the economy and sex discrimination with respect to wages and occupational segregation, which affect women economists along with all other. *CSWEP* feels that it is vital that such research provide the basis for its policy recommendations. It is *CSWEP*'s acceptance of the responsibility to collect and analyze data relevant to the status of women economists and to further the theoretical and applied research related to the status of women in general that most sets the work of this Committee apart from that of caucuses in some other professional as-

[1] In addition to presidents of the *AEA* who served *ex officio*, membership from March 1972 to date has included Walter Adams, Michigan State University, Carolyn Shaw Bell, Wellesley College (Chair, 1972 and 1973), Francine Blau, University of Illinois, Martha Blaxall, Health, Education & Welfare, Washington, DC, Kenneth E. Boulding, University of Colorado, Janice Madden, University of Pennsylvania, Collette Moser, Michigan State University, Barbara B. Reagan, Southern Methodist University (Chair, 1974–76), Isabel Sawhill, Urban Institute, Washington, DC, Margaret Simms, Atlanta University; Myra Strober, Stanford University, Nancy Teeters, Budget Committee, House of Representatives, Washington, DC, Phyllis Wallace, Sloan School, Massachusetts Institute of Technology, Florence Weiss, National Economic Research Associates, New York City

sociations. Many of the specific activities of the Committee discussed below support more than one of the above general goals.

I. Roster

From the beginning, compilation of a roster of women economists with their fields and professional qualifications has been a major project of *CSWEP*. *AEA* membership lists had no sex identification when the Committee was organized, and even the subsequent *Directory of Members* in 1974, which did include sex in the basic coding of the data at the request of this Committee, gets out of date quickly. (Such a list is invaluable as a benchmark.) *CSWEP*'s roster of women economists has grown from the 300 or so entries at the end of our first year to well over 1,800 this year. The roster has been computerized for several years now, and in the fall of 1976 we again updated our file by sending each woman economist listed a copy of the material she previously supplied us and asking for the most recent information in areas of specialization, highest degree in economics, professional grade or rank, and address. This fall we added a question on availability for new employment even though we realize that availability depends on the nature of the job offer New names are added to the roster at the request of individual women. Sometimes this comes about because they hear about our work and write asking to be put on our list. Some have come through open meetings we have held from time to time in major cities such as New York, Chicago, Boston, and Washington, D.C. Many women economists come by the room we "women" at the *AEA* annual meeting, and when they register with us, they ask to be added to our roster. In addition, we have asked academic departments of economics to give us names of women economists employed. In the fall of 1976, to save time of the respondent, the question on the Universal Academic Questionnaire was limited to asking for the names of new employees hired who are women economists. Another year, the question will revert to the broader question asking for names of all women economists employed. *CSWEP* has an ongoing problem of identifying women economists in business and government and urges all our colleagues to help us find and register such women.

Prospective employers may contact *CSWEP* and, for a fee to help defray Committee costs, they may receive a list of women economists who meet criteria specified by the prospective employers such as highest degree in economics, number of years of experience, field, or present rank. They can then contact the women so listed to ask for more detailed data on their careers and ascertain their interest in particular jobs. The use of this service is growing.

II. *CSWEP* Newsletter

Quarterly *CSWEP* Newsletters have been sent out to all women economists on our roster. The newsletter gives our associate members news of committee activities, asks for their help in various *AEA* activities, lists requests for articles, conference and program plans and participants, grant and fellowship opportunities, and regional activities, and notes research findings or publications of possible interest *AEA* members who would like to submit short items for inclusion in the *CSWEP* Newsletter are encouraged to send them in to us in writing. The Newsletter clearly is an important way in which we have widened the informal network and reduced the sense of isolation felt by some. Numerous associate members have contributed $5.00 or more to help defray the costs of the Newsletter in appreciation of this service. Additional contributions are welcome.

The *CSWEP* Newsletter for the past five years has carried brief announcements of job openings for economists for any prospective employer who asks us in writing to do so. In the fall of 1976 a survey of a sample of employers for whom we have carried such notices is being conducted so as to help evaluate the usefulness of this service.

III. Annual Meeting

Each year since its founding, *CSWEP* has

taken responsibility for organizing one session of the program for the annual meeting of the *AEA*. We have used this opportunity to present findings of the Committee or to organize presentations and discussions of research papers related to discrimination, women in the labor market, or other economic aspects of women's roles. This year *CSWEP* organized a session on the economics of the two-earner family. The quality of the papers presented was high and interest in the topics was great.

At the Atlantic City meetings, as last year, a special room in the heart of the meeting room area was set up for *CSWEP*. The room was kept open throughout the meetings by committee members and volunteers from among our associate members It became an open seminar on the economic status of women economists, job opportunities, and general discussion of colleagues' concerns and *CSWEP* activities. The room was well used throughout the meetings and discussion was lively.

Because of the early date of the 1976 annual meetings, less than normal activity occurred at Atlantic City related to job opportunities. *CSWEP* will have informal supportive services for women economists at the follow-up Job Market Session planned by the *AEA* for early January in Chicago to meet the special needs for a job market this year.

IV. Research

Occupational Segregation. The nearly three-year effort of *CSWEP* funded by a special grant from the Carnegie Corporation of New York to analyze occupational segregation by sex, the forces behind it, and its implications for women's economic standing was brought to completion in 1976. The final results of this effort were published (see Blaxall and Reagan). *CSWEP* members designed, planned, and participated in the national workshop conference on occupational segregation in May 1975 on which the book is based. The conference was jointly sponsored by *CSWEP* and the Center for Research on Women in Higher Education and the Professions at Wellesley College. Papers for the conference were commissioned 1) to accompany economic analyses of labor market phenomena with analysis from other disciplines in the social sciences and discussion of that work by economists so as to analyze occupational segregation as an interlocking set of institutions with sociological, psychological, and economic aspects and with deep historical roots; and 2) to consider what policy changes might be needed to achieve a society free from denial of job opportunities on the basis of sex. This book fulfills *CSWEP*'s promise to disseminate the research findings. We again thank the Carnegie Corporation of New York for making this effort financially possible. Occupational segregation by sex and the forces behind it are major factors in the low proportion of women currently among economists. Furthermore, even within economics, *CSWEP* research has revealed that some fields of specialization have considerably higher proportions of women than other areas of specialization, thus demonstrating the pervasiveness of sexism in occupational segregation down to micro levels. *CSWEP* hopes that this research effort to better understand occupational segregation will help provide the understanding of basic forces that is needed to overcome occupational segregation in general, and in particular to overcome the view that economics is a more appropriate professional field for men than for women.

1974-1975 Survey of Economists Work continued this year in building a computer tape of the data from *CSWEP*'s survey of a paired sample of more than 1,200 male and female economists who did their academic work for their highest degree at the same time in the same university. Data collection was completed in 1975. Preliminary results from the first 710 women respondents and the first male matches received were published in three articles in the May 1975 *American Economic Review* (Alice Amsden and Moser, Strober, and Reagan). The major preliminary finding was that salaries of women economists are substantially below those of their male counterparts even when their educational attainments and work histories are simi-

lar. More refined analysis of the complete sample, particularly the returns from economists with Ph.D.s, is continuing with emphasis on reasons for salary differences between men and women, and a search for differences in career patterns since obtaining Ph.D.s in economics.

Academic Labor Market for Economists, by Sex. Starting in the fall of 1972, *CSWEP* undertook a survey of departments of economics and agricultural economics in colleges and universities to obtain more information on the demand for economists, promotion rates, and production of graduate and undergraduate majors in economics, all classified by sex. This project, on recommendation of *CSWEP*, grew in 1975 into the Universal Academic Questionaire administered by the Nashville office of the *AEA*. *CSWEP* has continued to analyze the resulting data by sex. In previous years, such analysis has been published in the May issue of the *American Economic Review* in the annual report of *CSWEP*. This year our analysis of the 1976 data will be published in the Notes section of the June 1977 issue of the *American Economic Review* because the survey had not been made at the time this report had to go to press because of the change in the time of the 1976 annual meeting of the *AEA*.

As a service to the *AEA* this year, *CSWEP* bore the costs of putting on the computer the results of the 1975 survey made by the *AEA* with the Universal Academic Questionaire.

BARBARA B. REAGAN, *Chairperson*

REFERENCES

Alice H. Amsden and Collette Moser, "Job search and Affirmative Action," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 83–91.

Martha Blaxall and Barbara B. Reagan, *Women and the Workplace, The Implications of Occupational Segregation*, Chicago 1976.

Kenneth E. Boulding and ———, "Combatting Role Prejudice and Sex Discrimination," *Amer. Econ. Rev.*, Dec. 1973, *63*, 1049–61.

Barbara B. Reagan, "Two Supply Curves for Economists? Implications of Mobility and Career Attachment of Women," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 100–07.

———, "Report of the Committee on the Status of Women in the Economics Profession," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 512–20.

Myra H. Strober, "Women Economists: Career Aspirations, Education, and Training," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 92–99.

Report of the Representative to the National Bureau of Economic Research

National Bureau of Economic Research (NBER) Programs. Staff projects in 1975 and 1976 were in these areas: computer algorithms for applied economic and management research, including mathematical programming, exploratory data analysis, modeling, and econometrics; measurement of economic and social performance, including national accounting, price indexes, productivity, employment, inflation, business cycle indicators, international economic indicators, and forecasting; financial and industrial institutions and processes; urban and regional studies; international economic relations; trade flows and adjustments; and human behavior and social institutions, including factors affecting income distribution, law and economics, health and health care, education, population and family economics. Offices are in New York, Washington, Palo Alto and Cambridge.

Conference and Workshops. Conferences on Research in Income and Wealth in 1975 dealt with The Economics of Residential Location and Urban Housing Markets (Gregory K. Ingram, Chairman), and with New Developments in Productivity Measurement (Beatrice N. Vaccara and John Kendrick, Chairpersons). At press time a conference on Measurement of Capital and Related Aggregates was planned for October 1976 (Dan Usher, Chairman). A conference on wealth is planned for 1977 (James D. Smith, Chairman).

The Universities-National Bureau Committee for Economic Research held a conference in 1975 on the Economic Analysis of Political Behavior (Melvin W. Reder, Chairman). Future conferences were planned at press time on Causes and Economic Effects of Population Changes in Less Developed Countries in October 1976 (Richard Easterlin, Chairman) and on Low Income Labor Markets in 1977 (Sherwin Rosen, Chairman). For future conferences, three exploratory committees were considering Taxation and Household Behavior (Martin Feldstein, Chairman), Economic Planning and Regulation (Richard Nelson, Chairman), and Economics of Information and Uncertainty (George Stigler, Chairman).

The Conference on Econometrics and Mathematical Economics operates through ongoing working groups centered at host universities. In 1975 and 1976 over two dozen seminars were held on topics including Comparison of Econometric Models, General Equilibrium Models, Monetary and Fiscal Analysis, Natural Resources and Economic Growth, and Bayesian Inference in Econometrics.

The *NBER* and the Bureau of the Census were co-sponsors of a conference on Seasonal Analysis of Economic Time Series held in September, 1976.

The Conference on the Computer in Economic and Social Research held six workshop sessions during 1975 and 1976 on Estate Multiplier Estimates of Wealth, Pension Research, Stochastic Control, Time Use, Microanalytic Models, and Current Research in the Economics of Education. Another meeting is scheduled for May 25–27, 1977, in New Haven, Connecticut, with Ray Fair, David Kendrick and Edison Tse as coordinators. Its purpose is to bring together economists and engineers to share ideas about 1) application of control theory methods to a wide range of economic problems, and 2) improvements in the optimization and estimation methods used by control theorists.

The Latin American Computer Workshop program included conferences on Monetary Correction or Indexation and on Planning and Short-Term Macroeconomic Policy in Latin America. A conference is now planned for early February 1977, in Bogota, Colombia, on Commodity Markets, Models and Policies in Latin America.

Under the *U.S. - USSR* Scientific and Tech-

nical Program of Cooperation in the Field of Application of Computers to Management, the *NBER* has continued to help plan and coordinate a program of cooperation in econometric models and modeling of large-scale systems The *U.S.* participation is financed by the National Science Foundation. Exchange visits have been made by delegations of researchers and experts from the two countries, and conferences have been held on econometric modeling methods and the use of computers in management information and operating systems. John R. Meyer and Harvey J McMains have been the coordinators for the *NBER*'s part of the program.

Fellowships. For 1975–76 *NBER* Faculty Research Fellowships were awarded to Donald O. Parsons of Ohio State University and David L. Rubinfeld of the University of Michigan For 1976–77 the Faculty Research Fellows are Daniel A. Graham of Duke University, Cheng Hsiao of the University of California, Berkeley, and J. Huston McCulloch of Boston College.

Foreign Research Fellowships at the *NBER* were awarded for 1975–76 to Patricio Meller, Catholic University of Chile, Romeo M. Bautista, University of the Philippines, and Oey Astra Meesook, Thammasat University, Bangkok. For 1976–77 Narongchai Akrasanee, also of Thammasat University, is a Foreign Research Fellow.

Publications. In 1975 the *NBER* published thirteen books, including three conference volumes and ten staff research reports. In 1976 the *NBER* has scheduled the publication of ten books, including four conference volumes and six staff research reports. The *NBER* also published a quarterly journal of research papers, *Explorations in Economic Research*, and the quarterly *Annals of Economic and Social Research*, a journal dealing with computers, information retrieval, and research methodology.

The *Annual Report* of the *NBER* contains more information about officers, directors, staff, the research program, and publications. It is available on request to the *NBER*, 261 Madison Avenue, New York, NY 10016.

Officers. In April, 1976, John R. Meyer announced his intention of retiring from the Presidency of the *NBER*, as of the annual meeting in September 1977, and a committee of directors has been appointed to conduct a search and recommend a person to be his successor.

Officers elected for the ensuing year at the September 1976 annual meeting of the *NBER*'s Board of Directors included J. Wilson Newman, Chairman; Moses Abramovitz and James J. O'Leary, Vice Chairmen; and John R. Meyer, President.

CARL F. CHRIST, *Representative*

Report of the Committee on U.S.-Soviet Exchanges

After preliminary discussions extending over a couple of years, the Executive Committee in 1974 authorized an invitation to a group of Soviet economists to visit the United States in the spring of 1975. To organize this visit and possible subsequent exchanges, it created a Committee on U.S.-Soviet Exchanges, consisting of Abram Bergson, John Meyer, and Fritz Machlup (Chairman). The sponsoring body in the Soviet Union is the Association of Soviet Scientific Economic Institutions, under the chairmanship of Academician Tigran Khachaturov.

Pursuant to this invitation, ten Soviet economists came to the United States for two weeks in April, 1975. There were several days of seminar discussion in Washington with *U.S.* economists, on problems of the effectiveness of capital investment and economic relations between the two countries. These discussions suffered somewhat from lack of preparation on both sides. The Soviet papers were not available for distribution before the meeting. The *U.S.* participants were assembled on rather short notice and were a rotating group, since the meeting occurred during the university term when people could get away only for short periods.

After the Washington discussions, the Soviet group visited universities and research centers in Philadelphia, Princeton, New York, and Cambridge, Mass. Living costs and travel expenses within the United States were covered by the American Economic Association *(AEA)* under a grant from *IREX*, while the *USSR* paid for travel to and from the United States.

In the fall of 1975 the Soviet Association issued a return invitation for ten American economists to visit the *USSR* in June 1976, for a conference in Moscow on "The Economics of Technological Progress" and for visits to economic institutes in other parts of the country. Meanwhile, Fritz Machlup had asked to be relieved of the Chairmanship of the *AEA* Committee, and Lloyd Reynolds had been asked to succeed him, with Abram Bergson and John Meyer continuing to serve as members.

With longer lead time, it was possible to make more systematic preparations than those for the 1975 meeting. Most of the *U.S.* participants were invited in December, 1975, though a few later replacements were necessary because of schedule conflicts. Abstracts of the *U.S.* papers were sent to Moscow in May, both in English and in Russian translation. Abstracts of the Soviet papers, in English translation, arrived in the United States in May, in time to be reproduced for distribution to the *U.S.* participants. A request for funds, mainly to cover international air travel costs, was submitted to and eventually approved by the Office of International Programs of the National Science Foundation. Travel and living costs within the Soviet Union were covered by the host organization, following the procedure established in 1975.

The members of the *U.S.* group included: Abram Bergson, Edward F. Denison, Evsey Domar, Rendigs Fels, John W. Kendrick, M. Ishaq Nadiri, Merton J. Peck, Lloyd G. Reynolds, Frederick Scherer and Vladimir Treml.

I. Report on the June, 1976 Exchange

The first four days after arrival were spent in Moscow for round-table discussion of the prepared papers. Instead of reading papers, each participant was asked to summarize his main points in ten or fifteen minutes, leaving most of the time free for question and answer discussion. Simultaneous translation, on the whole quite satisfactory, was provided by two interpreters. In addition to those giving papers, the meetings were attended by Soviet economists from a number of research institutes, including the Institute of Economics, the Institute of World Economics and International Relations, the Central Economic-Mathematics Institute, and the Institute of the U.S.A. and Canada. The papers given at the meeting are being compiled in a mimeographed volume, which will be available to interested scholars in a few months' time.

The round-table discussion was franker and

more interesting than the papers themselves. The Soviet participants asked intelligent questions and made significant comments. There was a pleasing absence of *pro forma* ideological pronouncements. The atmosphere was genuinely friendly, and not very different from what one would encounter at conferences elsewhere in the world.

In addition to the conference sessions, the *U.S.* group visited the Faculty of Economics at Moscow State University, the leading center in the country for university training and research in economics; and the Central Economics-Mathematics Institute, for discussion of its research program with Academician Federenko and senior members of his staff

From Moscow we proceeded to Kiev, where our host was the Institute of Economics of the Ukrainian Academy of Sciences. This is a sizeable organization, with a professional staff of about 150 economists, and with considerable specialization in the economy of the Ukrainian Republic. We spent one morning discussing their research program with the Director and senior staff members On another morning three members of the *U.S* delegation lectured to the Institute staff and answered questions, which were numerous and lively.

The final stop was in Leningrad. Here we visited the Institute of Economic and Social Problems, a large, relatively new, multidisciplinary research organization with considerable emphasis on urban problems; and the Voznesensky Institute, a very large teaching and research organization, which is a major supplier of economists for banks, state enterprises, and planning bodies. They seem to regard themselves as the Soviet equivalent of Harvard Business School. The curriculum includes considerable emphasis on quantitative techniques and all students receive some computer training.

The results of such a visit are intangible and difficult to evaluate. The Soviet participants seemed keen to be in touch with Western economics and economists. On several occasions there was mention of the possibility of joint research by Soviet and American economists, though the question of how this might be organized was not pursued. The younger Soviet economists, in particular, seemed well-read in the Western literature, including the writings of some of the *U.S.* delegation. We had occasion to clear up a number of technical issues, and even to dispense thesis advice to a younger Ph.D. candidate working on inflation in the Western countries. The Western approach to research, and particularly the importance we attach to measurement, may have had some useful demonstration effect on the Soviet participants.

As for the *U.S.* participants, one cannot say that we learned much that is new about the Soviet *economy*. But we did learn a good deal about how economic research is organized, what researchers are doing, the content of university curricula in economics, the labor market for economics graduates, and related matters. The personal contacts established during the visit should also be useful in securing cooperation for *U.S.* graduate students and others interested in doing economic research in the Soviet Union.

II. Future Plans

The two members of our Committee (Bergson and Reynolds) who were in Moscow took the opportunity for tentative discussion of future plans with Academician Khachaturov. There is clearly a stong interest on the Soviet side in continuing the exchange arrangement. A half-dozen possible themes for the next meeting were discussed at some length. Among these, Khachaturov expressed some preference for an area comprising worker training, deployment, compensation, and productivity. Tentative plans are underway for a June 1977 conference in this area.

The Executive Committee should consider, we think, whether there is sufficient payoff from the program to warrant its continuation on an experimental basis. An affirmative answer would imply extension for another two-year round of exchanges, i.e., a Soviet visit to the United States in 1977, followed by a *U.S.* visit to the *USSR* in 1978. The National Science Foundation is a possible, though not an assured, source of funding for this purpose.

LLOYD G. REYNOLDS, *Chairperson*

Report of the *Ad Hoc* Committee On Federal Funding of Economic Research

Should the American Economic Association, (*AEA*) seek to stimulate economic research? A majority of the Committee believes that is should, proposing below two main directions of endeavor. A minority believes that such action will involve the Association in a lobbying, political, process—unwisely so where effective, and doubly unwise where ineffective.

I. The Determinants of Economic Research

The amount of economic research done in the United States, as its direction, is determined largely by: the extent to which economists allocate their own time and money to research; the judgments of Congressmen and bureau chiefs; and the goals of foundation and business executives. The *U.S.* government provides the largest source of basic research funding. That allocation, in turn, is shaped by the endeavors and the ingenuity which the members of other scholarly disciplines display—as the distribution e.g., of National Science Foundation (*NSF*) and *ERDA* expenditures by discipline suggests.

A majority of the Committee is persuaded that the above process induces less than the optimum amount of economic research.[1] Some knowledgeable economists who wrote us have indicated their belief that the occasional whims of a single Congressman have arbitrarily cut *NSF* funding of economic research. At the same time individual committee members felt that certain major projects in other federal agencies were of little, or negative, value. In these varied instances the choice of projects was not determined by peer review functioning within budget guidelines.

II. Increasing Federal Expenditures on Economic Research

A majority of the Committee holds that Federal expenditures 1) on basic economic research, and 2) on basic data collection, can usefully be expanded. It believes that *AEA* support of such expansion is both appropriate and desirable. It does not propose any system of priority nor particular modes by which such support should be provided, leaving those to be developed by the Executive Committee.[2] It premises the importance of such activities for the useful growth of the discipline.

Most specifically the majority believes that *NSF* funding for economic research (Division of Social Sciences) could be expanded by, say, 20 to 30 percent with no decline in the quality and usefulness of the projects to be approved.[3] That persuasion rests on individual contacts with the work of the Division (as advisers, panel members, project referees), as well as on the finding in the recent Simon committee report to the National Academy of Sciences-National Research Council.

There was, however, no corresponding support for the *AEA* to become involved in supporting "applied research." We received some comment on work hitherto done for *NSF-Research Applied to National Needs*, as well as on some fairly substantial programs sponsored by individual cabinet departments and agencies. The development of such research, it seemed to be agreed, should be left to the individual entrepreneurs (academic and other) and the agencies seeking solutions to problems.

[1]A minority is not persuaded we could identify such an optimum if we saw it.

[2]One minority view finds that if support is to have real impact it must pick and choose among the host of budgetary items that fall under the imprecise rubric of "basic economic research." Yet doing so would involve the *AEA* unduly, and improperly, in both representation and administration, and thereby in lobbying—an improper, imprudent, and probably ineffectual activity for a scientific organization

Another minority view finds that offering "general support" is a vague and ineffectual counsel It argues that the *AEA* should 1) open a Washington office, and 2) follow procedures something like those already pursued by other "scientific organizations," who do so both effectively and without apparent objection.

[3]The majority notes, as well, that particular members of Congress, or columnists may fasten on particular projects, and end by cutting back on basic research, yet no agency is in a position to argue the opposite position, while few individuals will care to do so

The committee did not address itself to the question of achieving an optimum amount, or allocation, of economic research. For example, more (and better?) economic research might perhaps be achieved by requiring full professors to continue publication in order to retain their status. Perhaps better data for economic research could be provided by cutting out some major federal data collection projects and then shifting part of such expenditures to more useful research. The committee members were not young enough to solve such questions before the natural term of their days, or to believe that they could. The majority therefore settled for their best leadings, based on insights to date.

III. Directions of Further Inquiry

The Committee had neither the time nor the inclination to conduct inquiries beyond those involved in establishing its primary judgments on the central issue.[4] However it recognizes certain inquiries as relevant, particularly if the Executive Committee decides for a more active role in supporting economic research. Such inquiries may well include the following.

(1) Both the Council of Economic Advisers and the Federal Reserve Board (FRB) are centrally important agencies that rely on economic research. How seriously and extensively do they try to develop basic research? or data collections that underly, e.g., the national income accounts, price indices, anticipation statistics, on which they regularly rely? What do their key staff members believe the role of the *AEA* should be?

(2) Congress decisively shapes the amount of funding for economic research. But we have almost no knowledge of how Congressmen decide these matters. What considerations impel key members of the Joint Economic Committee, the Joint Committee on Internal Revenue Taxation, Banking and Currency, when they come to vote on funding of research by the *NSF*, *FRB* or on data collection by Census, Bureau of Labor Statistics? Some inquiry among these legislators, and their staff, is desirable.

(3) Economists, though not the *AEA*, have previously made careful and extended reviews of data basic to much economic research. (Such studies include ones by the National Bureau of Economic Research on prices, national income; by *FRB* consultant committees on inventories, consumer intentions; etc.) What has been the result of all these recommendations? Study might give some insight into the more useful directions, and the likelihood of success, for exhortations by the *AEA*.

MILTON FRIEDMAN*
GARY FROMM
ZVI GRILICHES
ROBERT SOLOW
STANLEY LEBERGOTT, *Chairperson*

[4]A minority view would emphasize that the lack of such factual inquiries makes the Committee recommendations mostly general value judgments, which lack the simplest testing against experience. Such testing, however, might lead the *AEA* in quite a different direction

* *Has not seen the final report and cannot be held responsible*

Report of the Committee On Elections

In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee, the Electoral College, and the Committee on Elections.

The Nominating Committee, consisting of Kenneth Arrow, Chairperson, William James Adams, Duran Bell, Marianne Ferber, Jack Hirschleifer, Leonard Rapping, and Vincent Tarascio, submitted the nominations listed below for Vice Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-elect
Jacob Marschak

Vice-Presidents	*Executive Committee*
Robert Eisner	Marcus Alexis
Albert O. Hirschman	Samuel Bowles
Anne O. Krueger	Robert J Lampman
Roy Radner	Marc Nerlove

The Secretary prepared biographical sketches of the candidates and distributed ballots in late summer. The Committee on Elections, consisting of Ben Bolch, Chairperson, Barbara Haskew, and C. Elton Hinshaw, *ex officio*, canvassed the ballots and filed the following results:

Number of envelopes without names for identification	205
Number of envelopes received too late	73
Number of defective ballots	22
Number of legal ballots	4,692
	4,992

On the basis of the canvass of the votes, I certify that the following persons have been duly elected to the respective offices:

President-elect (for a term of one year)
Jacob Marschak

Vice-Presidents (for a term of one year)
Robert Eisner
Anne O. Krueger

Members of the Executive Committee (for a term of three years)
Robert J. Lampman
Marc Nerlove

In accordance with the actions of the Executive Committee at its meetings on December 27, 1975 and March 19, 1976, amendments to Article I, Section 2; Article I, Section 3; Article II; Article III, Sections 2, 3, and 4; Article IV, Sections 2, 3, 4, 5, 6, and 7; Article V, Section 1; and Article VI of the bylaws were submitted to the members in a mail ballot in conjunction with the balloting for officers. The ballots were canvassed by the Committee on Elections. On the basis of the canvass, I certify that the amendments were approved.

The bylaws as amended now read:

Article I, Section 2 and 3.

2. There shall be six classes of members other than honorary: regular members with the academic rank of assistant professor or lower or with annual incomes of $12,000 or less irrespective of rank, paying the base fee defined below; regular members with the academic rank of associate professor or with annual incomes above $12,000 but not more than $20,000 paying one and one fifth times the base fee; regular members with the academic rank of full professor or with annual incomes above $20,000, paying one and two fifths times the base fee; family members (two or more persons living at the same address, second membership without subscription to the publications of the Association) paying one-fifth of the base fee; junior members (available to registered students for three years only) paying one-half the base fee; and life members comprising those who qualified for life membership by making a single payment of the designated amount prior to January 1, 1976, and exempt from annual fees.

Effective January 1, 1976, the base fee is $25.00 per year. The Executive Committee may increase the dues schedule, including both the base fee and the income brackets for regular members, in proportion to the increase occurring after January 1, 1976 in relevant price and wage indexes, provided that the increase in any year shall not exceed ten percent.

3. Foreign economists of distinction may be elected honorary members of the Association. The Executive Committee is authorized to determine the number of foreigners to be elected honorary members. Past presidents of the Association and members who have been awarded the Walker Medal shall be Distinguished Fellows. In addition, the Executive Committee may elect additional Distinguished Fellows, but not more than two in any one calendar year, from economists of high distinction in the United States and Canada. Candidates for Distinguished Fellowships shall be nominated by the Nominating Committee or the Executive Committee, and they shall be elected by the combined vote of the two committees. The Nominating Committee shall solicit and give due consideration to the recommendations of the Committee on Honors and Awards. The Nominating Committee is free to make no nominations in any particular year. However, it is not limited as to the number of candidates it may nominate in any year. Election to Distinguished Fellowship does not preclude election to any office of the Association.

Article II.

The Board of Trustees shall be composed of the voting members of the Executive Committee

Article III, Sections 2 and 3.

2. The association shall have the following officers who shall be appointed by the Executive Committee: a Secretary, a Treasurer, a Managing Editor of the *American Economic Review*, a Managing Editor of the *Journal of Economic Literature*, and a Counsel. The terms of office of each of these officers shall be three calendar years.

3. The Executive Committee shall consist of the President, the President-elect, two Vice-Presidents, the Secretary, the Treasurer, the two Managing Editors, the two ex-Presidents who have last held office, and six elected members, provided the Secretary, the Treasurer, and the two Managing Editors shall not be entitled to vote in the Executive Committee's meeting.

Article IV, Sections 2, 3, 4, 5, 6, 7, and 8.

2 Before October 1 of each year, the President-elect of the Association shall appoint a Nominating Committee for the following year, this Committee to consist of a past officer as Chairman and not less than five other members of the Association. In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee.

The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee. The Nominating Committee for each year shall be instructed to present to the Executive Committee on or before March 31 a nominee for the President-elect and two or more nominations for each other elective office to be filled, except the presidency, all these nominees being members of the Association. The members of the Nominating and Executive Committees shall constitute an Electoral College which shall consider the nominee of the Nominating Committee for the President-elect and select a single candidate for that office In the voting in the Electoral College each member shall have one vote provided that the number of members of the Nominating Committee present does not exceed the number of members of the Executive Committee present; otherwise, the members of the Nominating Committee present shall have fractional votes such that their sum equals the number of members of the Executive Committee present.

The Secretary shall inform all members of the Association of the actions of the Nominating Committee and the Electoral College not later than the June issue of the *American Economic Review*. An additional nomination for any office may be made by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-elect and not less than 4 percent for each of the other offices No member of the Association may validly petition for more than one nominee for the Executive Committee, one nominee for Vice-President, and one nominee for President-elect.

The election of officers by the membership shall take place by a mail ballot conducted by the Secretary each year. The ballot shall list all nominees alphabetically with indication "nominated by petition" where applicable. Space

shall be provided on the ballot for the individual voter's alternative choice for all offices. The Secretary shall mail the ballots to all members as soon as practicable after August 1 and set a deadline for receipt of ballots in the Secretary's office no earlier than October 1 and no later than November 1. The candidates with the highest number of votes for the various offices will be elected. The results of the election shall be certified and announced by the Secretary at the annual business meeting or in the *American Economic Review*.

3. The President-elect shall be responsible for the program for the annual meeting of the year in which he serves. He may at his discretion appoint a Program Committee to assist him.

4. The Secretary shall keep the records of the Association and perform such other duties as the Executive Committee may assign to him.

5. The Treasurer shall receive and have the custody of the funds of the Association, subject to the rules of the Executive Committee.

6 The Executive Committee shall have the control and management of the funds of the corporation. It may fill vacancies in the list of officers, and may adopt any rules or regulations for the conduct of its business not inconsistent with this constitution or with rules adopted at the annual meetings. It shall act as a committee on time and place of meetings and perform such other duties as the Association shall delegate to it A quorum shall consist of five voting members.

7. The Managing Editors shall, with the advice and consent of the Executive Committee, appoint members of Editorial Boards to assist them. They shall be ex officiis members and chairpersons of their respective Boards, which shall have charge of the publication of the *American Economic Review* and the *Journal of Economic Literature*.

8. The office of the corporation for legal purposes shall be at the office of the Counsel in the District of Columbia, and legal process against the corporation may be served on said counsel.

Article V, Section 1.

1. The annual meeting of this corporation shall be held at such time and place as may be determined by the Executive Committee. Notice of such time and place shall be given by publication in the *American Economic Review*, at least three months prior to such meeting.

Article VI.

Amendments, after having been approved by a majority of the Executive Committee, may be adopted by a majority of votes cast in a mail ballot.

Ben Bolch, *Chairperson*

THE AMERICAN ECONOMIC REVIEW

March 1977

VOLUME 67, NUMBER 2

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members, $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

Articles

Shorter Papers

Number 78 of a series of photographs of past presidents of the Association



Franco Modigliani

The Monetarist Controversy or, Should We Forsake Stabilization Policies?

By FRANCO MODIGLIANI*

In recent years and especially since the onset of the current depression, the economics profession and the lay public have heard a great deal about the sharp conflict between "monetarists and Keynesians" or between "monetarists and fiscalists." The difference between the two "schools" is generally held to center on whether the money supply or fiscal variables are the major determinants of aggregate economic activity, and hence the most appropriate tool of stabilization policies.

My central theme is that this view is quite far from the truth, and that the issues involved are of far greater practical import. There are in reality no serious analytical disagreements between leading monetarists and leading nonmonetarists. Milton Friedman was once quoted as saying, "We are all Keynesians, now," and I am quite prepared to reciprocate that "we are all monetarists"—if by monetarism is meant assigning to the stock of money a major role in determining output and prices. Indeed, the list of those who have long been monetarists in this sense is quite extensive, including among other John Maynard Keynes as well as myself, as is attested by my 1944 and 1963 articles.

In reality the distinguishing feature of the monetarist school and the real issues of disagreement with nonmonetarists is not monetarism, but rather the role that should probably be assigned to stabilization policies. Nonmonetarists accept what I regard to be the fundamental practical message of *The General Theory*: that a private enterprise economy using an intangible money *needs* to be stabilized, *can* be stabilized, and therefore *should* be stabilized by appropriate monetary and fiscal policies. Monetarists by contrast take the view that there is no serious need to stabilize the economy; that even if there were a need, it could not be done, for stabilization policies would be more likely to increase than to decrease instability; and, at least some monetarists would, I believe, go so far as to hold that, even in the unlikely event that stabilization policies could on balance prove beneficial, the government should not be trusted with the necessary power.

What has led me to address this controversy is the recent spread of monetarism, both in a simplistic, superficial form and in the form of growing influence on the practical conduct of economic policy, which influence, I shall argue presently, has played at least some role in the economic upheavals of the last three years.

In what follows then, I propose first to review the main arguments bearing on the *need* for stabilization policies, that is, on the likely extent of instability in the absence of such policies, and then to examine the issue of the supposed destabilizing effect of pursuing stabilization policies. My main concern will be with instability generated by the traditional type of disturbances—demand shocks. But before I am through, I will give some consideration to the difficult problems raised by the newer type of disturbance—supply shocks.

I. The Keynesian Case for Stabilization Policies

A. *The General Theory*

Keynes' novel conclusion about the need for

*Presidential address delivered at the eighty-ninth meeting of the American Economic Association, Atlantic City, New Jersey, September 17, 1976. The list of those to whom I am indebted for contributing to shape the ideas expressed above is much too large to be included in this footnote. I do wish, however, to single out two lifetime collaborators to whom my debt is especially large, Albert Ando and Charles Holt. I also wish to express my thanks to Richard Cohn, Rudiger Dornbusch, and Benjamin Friedman for their valuable criticism of earlier drafts, and to David Modest for carrying out the simulations and other computations mentioned in the text.

stabilization policies, as was brought out by the early interpreters of *The General Theory* (for example, John Hicks, the author, 1944), resulted from the interaction of a basic contribution to traditional monetary theory—liquidity preference—and an unorthodox hypothesis about the working of the labor market—complete downward rigidity of wages.

Because of liquidity preference, a change in aggregate demand, which may be broadly defined as any event that results in a change in the market clearing or equilibrium rate of interest, will produce a corresponding change in the real demand for money or velocity of circulation, and hence in the real stock of money needed at full employment. As long as wages are perfectly flexible, even with a constant nominal supply, full employment could and would be maintained by a change of wages and prices as needed to produce the required change in the real money supply—though even in this case, stability of the price level would require a countercyclical monetary policy. But, under the Keynesian wage assumption the classical adjustment through prices can occur only in the case of an increased demand. In the case of a decline, instead, wage rigidity prevents the necessary increase in the real money supply and the concomitant required fall in interest rates. Hence, if the nominal money supply is constant, the initial equilibrium must give way to a new stable one, characterized by lower output and by an involuntary reduction in employment, so labeled because it does not result from a shift in notional demand and supply schedules in terms of real wages, but only from an insufficient real money supply. The nature of this equilibrium is elegantly captured by the Hicksian *IS-LM* paradigm, which to our generation of economists has become almost as familiar as the demand-supply paradigm was to earlier ones.

This analysis implied that a fixed money supply far from insuring approximate stability of prices and output, as held by the traditional view, would result in a rather unstable economy, alternating between periods of protracted unemployment and stagnation, and bursts of inflation. The extent of downward instability would depend in part on the size of the exogenous shocks to demand and in part on the strength of what may be called the Hicksian mechanism. By this I mean the extent to which a shift in *IS*, through its interaction with *LM*, results in some decline in interest rates and thus in a change in income which is smaller than the original shift. The stabilizing power of this mechanism is controlled by various parameters of the system. In particular, the economy will be more unstable the greater the interest elasticity of demand for money, and the smaller the interest responsiveness of aggregate demand. Finally, a large multiplier is also destabilizing in that it implies a larger shift in *IS* for a given shock.

However, the instability could be readily counteracted by appropriate stabilization policies. Monetary policy could change the nominal supply of money so as to *accommodate* the change in real demand resulting from shocks in aggregate demand. Fiscal policy, through expenditure and taxes, could *offset* these shocks, making full employment consistent with the initial nominal money stock. In general, both monetary and fiscal policies could be used in combination. But because of a perceived uncertainty in the response of demand to changes in interest rates, and because changes in interest rates through monetary policy could meet difficulties and substantial delays related to expectations (so-called liquidity traps), fiscal policy was regarded as having some advantages.

B. *The Early Keynesians*

The early disciples of the new Keynesian gospel, still haunted by memories of the Great Depression, frequently tended to outdo Keynes' pessimism about potential instability. Concern with liquidity traps fostered the view that the demand for money was highly interest elastic; failure to distinguish between the short- and long-run marginal propensity to save led to overestimating the long-run saving rate, thereby fostering concern with stagnation, and to underestimating the short-run propensity, thereby exaggerating the short-run multiplier. Interest

rates were supposed to affect, at best, the demand for long-lived fixed investments, and the interest elasticity was deemed to be low. Thus, shocks were believed to produce a large response. Finally, investment demand was seen as capriciously controlled by "animal spirits," thus providing an important source of shocks. All this justified calling for very active stabilization policies. Furthermore, since the very circumstances which produce a large response to demand shocks also produce a large response to *fiscal* and a small response to *monetary* actions, there was a tendency to focus on fiscal policy as the main tool to keep the economy at near full employment.

C. *The Phillips Curve*

In the two decades following *The General Theory*, there were a number of developments of the Keynesian system including dynamization of the model, the stress on taxes versus expenditures and the balanced budget multiplier, and the first attempts at estimating the critical parameters through econometric techniques and models. But for present purposes, the most important one was the uncovering of a "stable" statistical relation between the rate of change of wages and the rate of unemployment, which has since come to be known as the Phillips curve. This relation, and its generalization by Richard Lipsey to allow for the effect of recent inflation, won wide acceptance even before an analytical underpinning could be provided for it, in part because it could account for the "puzzling" experience of 1954 and 1958, when wages kept rising despite the substantial rise in unemployment. It also served to dispose of the rather sterile "cost push"–"demand pull" controversy.

In the following years, a good deal of attention went into developing theoretical foundations for the Phillips curve, in particular along the lines of search models (for example, Edmund Phelps et al.). This approach served to shed a new light on the nature of unemployment by tracing it in the first place to labor turnover and search time rather than to lack of jobs as such: in a sense unemployment is all frictional—at least in developed countries. At the same time it clarified how the availability of more jobs tends to reduce unemployment by increasing vacancies and thus reducing search time.

Acceptance of the Phillips curve relation implied some significant changes in the Keynesian framework which partly escaped notice until the subsequent monetarists' attacks. Since the rate of change of wages decreased smoothly with the rate of unemployment, there was no longer a unique Full Employment but rather a whole family of possible equilibrium rates, each associated with a different rate of inflation (and requiring, presumably, a different long-run growth of money). It also impaired the notion of a stable underemployment equilibrium. A fall in demand could still cause an initial rise in unemployment but this rise, by reducing the growth of wages, would eventually raise the real money supply, tending to return unemployment to the equilibrium rate consistent with the given long-run growth of money.

But at the practical level it did not lessen the case for counteracting lasting demand disturbances through stabilization policies rather than by relying on the slow process of wage adjustment to do the job, at the cost of protracted unemployment and instability of prices. Indeed, the realm of stabilization policies appeared to expand in the sense that the stabilization authority had the power of choosing the unemployment rate around which employment was to be stabilized, though it then had to accept the associated inflation. Finally, the dependence of wage changes also on past inflation forced recognition of a distinction between the short- and the long-run Phillips curve, the latter exhibiting the long-run equilibrium rate of inflation implied by a *maintained* unemployment rate. The fact that the long-run tradeoff between unemployment and inflation was necessarily less favorable than the short-run one, opened up new vistas of "enjoy-it-now, pay-later" policies, and even resulted in an entertaining literature on the political business cycle and how to stay in the saddle by riding the Phillips curve (see for example, Ray Fair, William Nordhaus).

II. The Monetarists' Attack

A. *The Stabilizing Power of the Hicksian Mechanism*

The monetarists' attack on Keynesianism was directed from the very beginning not at the Keynesian framework as such, but at whether it really implied a need for stabilization. It rested on a radically different empirical assessment of the value of the parameters controlling the stabilizing power of the Hicksian mechanism and of the magnitude and duration of response to shocks, given a stable money supply. And this different assessment in turn was felt to justify a radical downgrading of the *practical relevance* of the Keynesian framework as distinguished from its *analytical validity*.

Liquidity preference was a fine contribution to monetary theory but in practice the responsiveness of the demand for money, and hence of velocity, to interest rates, far from being unmanageably large, was so small that according to a well-known paper by Milton Friedman (1969), it could not even be detected empirically. On the other hand, the effect of interest rates on aggregate demand was large and by no means limited to the traditional fixed investments but quite pervasive. The difficulty of detecting it empirically resulted from focusing on a narrow range of measured market rates and from the fact that while the aggregate could be counted on to respond, the response of individual components might not be stable. Finally, Friedman's celebrated contribution to the theory of the consumption function (1957) (and my own work on the life cycle hypothesis with Richard Brumberg and others, reviewed by the author, 1975) implied a very high short-run marginal propensity to save in response to transient disturbances to income and hence a small short-run multiplier.

All this justified the conclusion that (i) though demand shocks might qualitatively work along the lines described by Keynes, quantitatively the Hicks mechanism is so strong that their impact would be *small* and *transient*, provided the stock of money was kept on a steady growth path; (ii) fiscal policy actions, like other demand shocks, would have *minor* and *transitory* effects on demand, while changes in money would produce *large* and *permanent* effects on money income; and, therefore, (iii) the observed instability of the economy, which was anyway proving moderate as the postwar period unfolded, was most likely the result of the unstable growth of money, be it due to misguided endeavors to stabilize income or to the pursuit of other targets, which were either irrelevant or, in the case of balance of payments goals, should have been made irrelevant by abandoning fixed exchanges.

B. *The Demise of Wage Rigidity and the Vertical Phillips Curve*

But the most serious challenge came in Friedman's 1968 Presidential Address, building on ideas independently put forth also by Phelps (1968). Its basic message was that, despite appearances, wages were in reality perfectly flexible and there was accordingly *no* involuntary unemployment. The evidence to the contrary, including the Phillips curve, was but a statistical illusion resulting from failure to differentiate between price changes and *unexpected* price changes.

Friedman starts out by reviving the Keynesian notion that, at any point of time, there exists a unique full-employment rate which he labels the "natural rate." An unanticipated fall in demand in Friedman's competitive world leads firms to reduce prices and also output and employment along the short-run marginal cost curve—unless the nominal wage declines together with prices. But workers, failing to judge correctly the current and prospective fall in prices, misinterpret the reduction of nominal wages as a cut in *real* wages. Hence, assuming a positively sloped supply function, they reduce the supply of labor. As a result, the effective real wage rises to the point where the resulting decline in the demand for labor matches the reduced supply. Thus, output falls not because of the decline in demand, but because of the entirely voluntary reduction in the supply of labor, in response to erroneous perceptions. Furthermore, the fall in employ-

ment can only be temporary, as expectations must soon catch up with the facts, at least in the absence of new shocks. The very same mechanism works in the case of an increase in demand, so that the responsiveness of wages and prices is the same on either side of the natural rate.

The upshot is that Friedman's model also implies a Phillips-type relation between inflation, employment or unemployment, and past inflation,—provided the latter variable is interpreted as a reasonable proxy for expected inflation. But it turns the standard explanation on its head: instead of (excess) employment causing inflation, it is (the unexpected component of) the rate of inflation that causes excess employment.

One very basic implication of Friedman's model is that the coefficient of price expectations should be precisely unity. This specification implies that whatever the shape of the short-run Phillips curve—a shape determined by the relation between expected and actual price changes, and by the elasticity of labor supply with respect to the perceived real wage—the long-run curve *must be vertical*

Friedman's novel twist provided a fresh prop for the claim that stabilization policies are not really needed, for, with wages flexible, except possibly for transient distortions, the Hicksian mechanism receives powerful reinforcement from changes in the real money supply. Similarly, the fact that full employment was a razor edge provided new support for the claim that stabilization policies were bound to prove destabilizing.

C. *The Macro Rational Expectations Revolution*

But the death blow to the already badly battered Keynesian position was to come only shortly thereafter by incorporating into Friedman's model the so-called rational expectation hypothesis, or *REH*. Put very roughly, this hypothesis, originally due to John Muth, states that rational economic agents will endeavor to form expectations of relevant future variables by making the most efficient use of all information provided by past history. It is a fundamental and fruitful contribution that has already found many important applications, for example, in connection with speculative markets, and as a basis for some thoughtful criticism by Robert Lucas (1976) of certain features of econometric models. What I am concerned with here is only its application to macro-economics, or *MREH*, associated with such authors as Lucas (1972), Thomas Sargent (1976), and Sargent and Neil Wallace (1976).

The basic ingredient of *MREH* is the postulate that the workers of Friedman's model hold rational expectations, which turns out to have a number of remarkable implications: (i) errors of price expectations, which are the only source of departure from the natural state, cannot be avoided but they can only be short-lived and random. In particular, there cannot be persistent unemployment above the natural rate for this would imply high serial correlation between the successive errors of expectation, which is inconsistent with rational expectations; (ii) any attempts to stabilize the economy by means of stated monetary or fiscal rules are bound to be totally ineffective because their effect will be fully discounted in rational expectations; (iii) nor can the government successfully pursue *ad hoc* measures to offset shocks. The private sector is already taking care of any anticipated shock; therefore government policy could conceivably help only if the government information was better than that of the public, which is impossible, by the very definition of rational expectations. Under these conditions, *ad hoc* stabilization policies are most likely to produce instead further destabilizing shocks.

These are clearly remarkable conclusions, and a major *re*discovery—for it had all been said 40 years ago by Keynes in a well-known passage of *The General Theory*:

> If, indeed, labour were always in a position to take action (and were to do so), whenever there was less than full employment, to reduce its money demands by concerted action to whatever point was required to make money so abundant rela-

> tively to the wage-unit that the rate of interest would fall to a level compatible with full employment, we should, in effect, have monetary management by the Trade Unions, aimed at full employment, instead of by the banking systems. [p. 267]

The only novelty is that *MREH* replaces Keynes' opening "if" with a "since."

If one accepts this little amendment, the case against stabilization policies is complete. The economy is inherently pretty stable—except possibly for the effect of government messing around. And to the extent that there is a small residual instability, it is beyond the power of human beings, let alone the government, to alleviate it.

III. How Valid Is the Monetarist Case?

A. *The Monetarist Model of Wage Price Behavior*

In setting out the counterattack it is convenient to start with the monetarists' model of price and wage behavior. Here one must distinguish between the model as such and a specific implication of that model, namely that the long-run Phillips curve is vertical, or, in substance, that, in the long run, money is neutral. That conclusion, by now, does not meet serious objection from nonmonetarists, at least as a first approximation.

But the proposition that other things equal, and given time enough, the economy will eventually adjust to any indefinitely maintained stock of money, or *n*th derivative thereof, can be derived from a variety of models and, in any event, is of very little practical relevance, as I will argue below. What is unacceptable, because inconsistent with both micro and macro evidence, is the specific monetarist model set out above and its implication that all unemployment is a voluntary, fleeting response to transitory misperceptions.

One may usefully begin with a criticism of the Macro Rational Expectations model and why Keynes' "if" should not be replaced by "since." At the logical level, Benjamin Friedman has called attention to the omission from *MREH* of an explicit learning model, and has suggested that, as a result, it can only be interpreted as a description not of short-run but of long-run equilibrium in which no agent would wish to recontract. But then the implications of *MREH* are clearly far from startling, and their policy relevance is almost nil. At the institutional level, Stanley Fischer has shown that the mere recognition of long-term contracts is sufficient to generate wage rigidity and a substantial scope for stabilization policies. But the most glaring flaw of *MREH* is its inconsistency with the evidence: if it were valid, deviations of unemployment from the natural rate would be small and transitory—in which case *The General Theory* would not have been written and neither would this paper. Sargent (1976) has attempted to remedy this fatal flaw by hypothesizing that the persistent and large fluctuations in unemployment reflect merely corresponding swings in the natural rate itself. In other words, what happened to the United States in the 1930's was a severe attack of contagious laziness! I can only say that, despite Sargent's ingenuity, neither I nor, I expect, most others at least of the nonmonetarists' persuasion are quite ready yet to turn over the field of economic fluctuations to the social psychologist!

Equally serious objections apply to Friedman's modeling of the commodity market as a perfectly competitive one—so that the real wage rate is continuously equated to the *short-run* marginal product of labor—and to his treatment of labor as a homogenous commodity traded in an auction market, so that, at the going wage, there never is any excess demand by firms or excess supply by workers. The inadequacies of this model as a useful formalization of present day Western economies are so numerous that only a few of the major ones can be mentioned here.

Friedman's view of unemployment as a voluntary reduction in labor supply could at best provide an explanation of variations in labor force—and then only under the questionable assumption that the supply function has a sig-

nificantly positive slope—but cannot readily account for changes in unemployment. Furthermore, it cannot be reconciled with the well-known fact that *rising* unemployment is accompanied by a fall, not by a *rise* in quits, nor with the role played by temporary layoffs to which Martin Feldstein has recently called attention. Again, his competitive model of the commodity market, accepted also in *The General Theory*, implies that changes in real wages, adjusted for long-run productivity trend, should be significantly negatively correlated with cyclical changes in employment and output and with changes in money wages. But as early as 1938, John Dunlop showed that this conclusion was rejected by some eighty years of British experience and his results have received some support in more recent tests of Ronald Bodkin for the United States and Canada. Similar tests of my own, using quarterly data, provide striking confirmation that for the last two decades from the end of the Korean War until 1973, the association of trend adjusted real compensations of the private nonfarm sector with either employment or the change in nominal compensation is prevailingly positive and very significantly so.[1]

This evidence can, instead, be accounted for by the oligopolistic pricing model—according to which price is determined by *long-run* minimum average cost up to a mark-up reflecting entry-preventing considerations (see the author, 1958)—coupled with some lags in the adjustment of prices to costs. This model implies that firms respond to a change in demand by endeavoring to adjust output and employment, without significant changes in prices relative to wages; and the resulting changes in available jobs have their initial impact not on wages but rather on unemployment by way of layoffs and recalls and through changes in the level of vacancies, and hence on the length of average search time.

If, in the process, vacancies rise above a critical level, or "natural rate," firms will endeavor to reduce them by outbidding each other, thereby raising the rate of change of wages. Thus, as long as jobs and vacancies remain above, and unemployment remains below, some critical level which might be labeled the "noninflationary rate" (see the author and Lucas Papademos, 1975), wages and prices will tend to accelerate. If, on the other hand, jobs fall below, and unemployment rises above, the noninflationary rate, firms finding that vacancies are less than optimal —in the limit the unemployed queuing outside the gate will fill them instantly—will have an incentive to reduce their relative wage offer. But in this case, in which too much labor is looking for too few jobs, the trend toward a sustained decline in the rate of growth of wages is likely to be even weaker than the corresponding acceleration when too many jobs are bidding for too few people. The main reason is the nonhomogeneity of labor. By far the largest and more valuable source of labor supply to a firm consists of those already employed who are not readily interchangeable with the unemployed and, in contrast with them, are concerned with protecting their earnings and not with reestablishing full employment. For these reasons, and because the first to quit are likely to be the best workers, a reduction of the labor force can, within limits, be accomplished more economically, not by reducing wages to generate enough quits, but by firing or, when possible, by layoffs which insure access to a trained labor force when demand recovers. More generally, the inducement to

[1]Thus, in a logarithmic regression of private nonfarm hourly compensation deflated by the private nonfarm deflator on output per man-hour, time, and private nonfarm employment, after correcting for first-order serial correlation, the latter variable has a coefficient of .17 and a t-ratio of 5. Similar though less significant results were found for manufacturing. If employment is replaced by the change in nominal compensation, its coefficient is .40 with a t-ratio of 6.5. Finally, if the change in compensation is replaced by the change in price, despite the negative bias from error of measurement of price, the coefficient of this variable is only −.09 with an entirely insignificant t-ratio of .7. The period after 1973 has been omitted from the tests as irrelevant for our purposes, since the inflation was driven primarily by an exogenous price shock rather than by excess demand. As a result of the shock, prices, and to some extent wages, rose rapidly while employment and real wages fell. Thus, the addition of the last two years tends to increase spuriously the positive association between real wages and employment, and to decrease that between real wages and the change in nominal wages or prices.

reduce relative wages to eliminate the excess supply is moderated by the effect that such a reduction would have on quits and costly turnover, even when the resulting vacancies can be readily filled from the ranks of the unemployed. Equally relevant are the consequences in terms of loss of morale and good will, in part for reasons which have been elaborated by the literature on implicit contracts (see Robert Gordon). Thus, while there will be some tendency for the rate of change of wages to fall, the more so the larger the unemployment—at least in an economy like the United States where there are no overpowering centralized unions—that tendency is severely damped.

And whether, given an unemployment rate significantly and persistently above the noninflationary level, the rate of change of wages would, eventually, tend to turn negative and decline without bound or whether it would tend to an asymptote is a question that I doubt the empirical evidence will ever answer. The one experiment we have had—the Great Depression—suggests the answer is negative, and while I admit that, for a variety of reasons, that evidence is muddied, I hope that we will never have the opportunity for a second, clean experiment.

In any event, what is really important for practical purposes is not the long-run equilibrium relation as such, but the speed with which it is approached. Both the model sketched out and the empirical evidence suggest that the process of acceleration or deceleration of wages when unemployment differs from the noninflationary rate will have more nearly the character of a crawl than of a gallop. It will suffice to recall in this connection that there was excess demand pressure in the United States at least from 1965 to mid-1970, and during that period the growth of inflation was from some 1.5 to only about 5.5 percent per year. And the response to the excess supply pressure from mid-1970 to early 1973, and from late 1974 to date was equally sluggish.

B. *The Power of Self-Stabilizing Mechanisms: The Evidence from Econometric Models*

There remains to consider the monetarists' initial criticism of Keynesianism, to wit, that even without high wage flexibility, the system's response to demand shocks is small and short-lived, thanks to the power of the Hicksian mechanism. Here it must be acknowledged that every one of the monetarists' criticisms of early, simpleminded Keynesianism has proved in considerable measure correct.

With regard to the interest elasticity of demand for money, post-Keynesian developments in the theory of money, and in particular, the theoretical contributions of William Baumol, James Tobin, Merton Miller, and Daniel Orr, point to a modest value of around one-half to one-third, and empirical studies (see for example, Stephen Goldfeld) are largely consistent with this prediction (at least until 1975!). Similarly, the dependence of consumption on long-run, or life cycle, income and on wealth, together with the high marginal tax rates of the postwar period, especially the corporate tax, and leakages through imports, lead to a rather low estimate of the multiplier.

Last but not least, both theoretical and empirical work, reflected in part in econometric models, have largely vindicated the monetarist contention that interest effects on demand are pervasive and substantial. Thus, in the construction and estimation of the MIT-Penn-Social Science Research Council (*MPS*) econometric model of the United States, we found evidence of effects, at least modest, on nearly every component of aggregate demand. One response to money supply changes that is especially important in the *MPS*, if somewhat controversial, is via interest rates on the market value of all assets and thus on consumption.

There is, therefore, substantial agreement that in the United States the Hicksian mechanism is fairly effective in limiting the effect of shocks, and that the response of wages and prices to excess demand or supply will also work *gradually* toward eliminating largely, if not totally, any effect on employment. But in the view of nonmonetarists, the evidence overwhelmingly supports the conclusion that the *interim* response is still of significant magnitude and of considerable duration, basically because the wheels of the offsetting mechanism grind slowly. To be sure, the first link of the mechanism, the rise in short-term rates, gets promptly into play and

heftily, given the low money demand elasticity; but most expenditures depend on long-term rates, which generally respond but gradually, and the demand response is generally also gradual. Furthermore, while this response is building up, multiplier and accelerator mechanisms work toward amplifying the shock. Finally, the classical mechanism—the change in real money supply through prices—has an even longer lag because of the sluggish response of wages to excess demand.

These interferences are supported by simulations with econometric models like the *MPS*. Isolating, first, the working of the Hicksian mechanism by holding prices constant, we find that a 1 percent demand shock, say a rise in real exports, produces an impact effect on aggregate output which is barely more than 1 percent, rises to a peak of only about 2 percent a year later, and then declines slowly toward a level somewhat over 1.5 percent.

Taking into account the wage price mechanism hardly changes the picture for the first year because of its inertia Thereafter, however, it becomes increasingly effective so that a year later the real response is back at the impact level, and by the end of the third year the shock has been fully offset (thereafter output oscillates around zero in a damped fashion). Money income, on the other hand, reaches a peak of over 2.5, and then only by the middle of the second year. It declines thereafter, and tends eventually to oscillate around a *positive* value because normally, a demand shock requires eventually a change in interest rates and hence in velocity and money income.

These results, which are broadly confirmed by other econometric models, certainly do not support the view of a highly unstable economy in which fiscal policy has powerful and everlasting effects. But neither do they support the monetarist view of a highly stable economy in which shocks hardly make a ripple and the effects of fiscal policy are puny and fast vanishing.

C. *The Monetarist Evidence and the St. Louis Quandary*

Monetarists, however, have generally been inclined to question this evidence. They countered at first with tests bearing on the stability of velocity and the insignificance of the multiplier, which, however, as indicated in my criticism with Albert Ando (1965), must be regarded as close to worthless. More recently, several authors at the Federal Reserve Bank of St. Louis (Leonall Andersen, Keith Carlson, Jerry Lee Jordan) have suggested that instead of deriving multipliers from the analytical or numerical solution of an econometric model involving a large number of equations, any one of which may be questioned, they should be estimated directly through "reduced form" equations by relating the change in income to current and lagged changes in some appropriate measure of the money supply and of fiscal impulses.

The results of the original test, using the current and but four lagged values of M^1 and of high Employment Federal Expenditure as measures of monetary and fiscal impulses, turned out to be such as to fill a monetarist's heart with joy. The contribution of money, not only current but also lagged, was large and the coefficients implied a not unreasonable effect of the order of magnitude of the velocity of circulation, though somewhat higher. On the other hand, the estimated coefficients of the fiscal variables seemed to support fully the monetarists' claim that their impact was both small and fleeting: the effect peaked in but two quarters and was only around one, and disappeared totally by the fourth quarter following the change.

These results were immediately attacked on the ground that the authors had used the wrong measure of monetary and fiscal actions, and it was shown that the outcome was somewhat sensitive to alternative measures; however, the basic nature of the results did not change, at least qualitatively. In particular, the outcome does not differ materially, at least for the original period up to 1969, if one replaces high employment outlays with a variable that might be deemed more suitable, like government expenditure on goods and services, plus exports.

These results must be acknowledged as disturbing for nonmonetarists, for there is little question that movements in government purchases and exports are a major source of demand disturbances; if econometric model estimates of

the response to demand disturbances are roughly valid, how can they be so grossly inconsistent with the reduced form estimates?

Attempts at reconciling the two have taken several directions, which are reviewed in an article coauthored with Ando (1976). Our main conclusion, based on simulation techniques, is that when income is subject to substantial shocks from many sources other than monetary and fiscal, so that these variables account for only a moderate portion of the variations in income (in the United States, it has been of the order of one-half to two-thirds), then the St. Louis reduced form method yields highly unstable and unreliable estimates of the true structure of the system generating the data.

The crucial role of unreliability and instability has since been confirmed in more recent work of Daniel O'Neill in his forthcoming thesis. He shows in the first place that different methods of estimation yield widely different estimates, including many which clearly overstate the expenditure and understate the money multipliers. He further points out that, given the unreliability of the estimates resulting from multicollinearity and large residual variance, the relevant question to ask is not whether these estimates differ from those obtained by structural estimation, but whether the *difference is statistically significant*; that is, larger than could be reasonably accounted for by sampling fluctuations.

I have carried out this standard statistical test using as true response coefficients those generated by the *MPS* model quoted earlier.[2] I find that, at least when the test is based on the largest possible sample—the entire post-Korean period up to the last two very disturbed years—the difference is totally insignificant when estimation is in level form (F is less than one) and is still not significant at the 5 percent level, when in first differences.

This test resolves the puzzle by showing that there really is no puzzle: the two alternative estimates of the expenditure multipliers are not inconsistent, given the margin of error of the estimates. It implies that one should accept whichever of the two estimates is produced by a more reliable and stable method, and is generally more sensible. To me, those criteria call, without question, for adopting the econometric model estimates. But should there be still some lingering doubt about this choice, I am happy to be able to report the results of one final test which I believe should dispose of the reduced form estimates—at least for a while. Suppose the St Louis estimates of the expenditure multiplier are closer to God's truth than the estimates derived through econometric models. Then it should be the case that if one uses their coefficients to forecast income beyond the period of fit, these forecasts should be appreciably better than those obtained from a forecasting equation in which the coefficients of the expenditure variable are set equal to those obtained from econometric models.

I have carried out this test, comparing a reduced form equation fitted to the period originally used at St. Louis, terminating in 1969 (but reestimated with the lastest revised data) with an equation in which the coefficients of government expenditure plus exports were constrained to be those estimated from the *MPS*, used in the above F-test. The results are clear cut: the errors using the reduced form coefficient are not smaller but on the average substantially *larger* than those using *MPS* multipliers. For the first four years, terminating at the end of 1973, the St. Louis equation produces errors which are distinctly larger in eight quarters, and smaller in but three, and its squared error is one-third larger. For the last two years of turmoil, both equations perform miserably, though even here the *MPS* coefficients perform just a bit better. I have repeated this test with equations estimated through the first half of the postwar period, and the results are, if anything, even more one-sided.

The moral of the story is pretty clear. First,

[2]For the purpose of the test, coefficients were scaled down by one-third to allow for certain major biases in measured government expenditure for present purposes (mainly the treatment of military procurement on a delivery rather than work progress basis, and the inclusion of direct military expenditure abroad)

reduced form equations relying on just two exogenous variables are very unreliable for the purpose of estimating structure, nor are they particularly accurate for forecasting, though per dollar of research expenditure they are surprisingly good. Second, if the St. Louis people want to go on using this method and wish to secure the best possible forecast, then they should ask the *MPS* or any other large econometric model what coefficients they should use for government expenditure, rather than trying to estimate them by their unreliable method.

From the theory and evidence reviewed, we must then conclude that opting for a constant rate of growth of the nominal money supply can result in a stable economy only in the absence of significant exogenous shocks. But obviously the economy has been and will continue to be exposed to many significant shocks, coming from such things as war and peace, and other large changes in government expenditure, foreign trade, agriculture, technological progress, population shifts, and what not. The clearest evidence on the importance of such shocks is provided by our postwar record with its six recessions.

IV. The Record of Stabilization Policies: Stabilizing or Destabilizing

A. *Was Postwar Instability Due to Unstable Money Growth?*

At this point, of course, monetarists will object that, over the postwar period, we have *not* had a constant money growth policy and will hint that the observed instability can largely be traced to the instability of money. The only way of meeting this objection squarely would be, of course, to rerun history with a good computer capable of calculating 3 percent at the helm of the Fed.

A more feasible, if less conclusive approach might be to look for some extended periods in which the money supply grew fairly smoothly and see how the economy fared. Combing through our post-Korean War history, I have been able to find just two stretches of several years in which the growth of the money stock was relatively stable, whether one chooses to measure stability in terms of percentage deviations from a constant growth or of dispersion of four-quarter changes. It may surprise some that one such stretch occurred quite recently and consists of the period of nearly four years beginning in the first quarter of 1971 (see the author and Papademos, 1976). During this period, the average growth was quite large, some 7 percent, but it was relatively smooth, generally well within the 6 to 8 percent band. The average deviation from the mean is about .75 percent. The other such period lasted from the beginning of 1953 to the first half of 1957, again a stretch of roughly four years. In sharp contrast to the most recent period, the average growth here is quite modest, only about 2 percent; but again, most four-quarter changes fell well within a band of two percentage points, and the average deviation is again .7. By contrast, during the remaining 13-year stretch from mid-1957 to the end of 1970, the variability of money growth was roughly twice as large if measured by the average deviation of four quarter changes, and some five times larger if measured by the percentage deviation of the money stock from a constant growth trend.

How did the economy fare in the two periods of relatively stable money growth? It is common knowledge that the period from 1971 to 1974, or from 1972 to 1975 if we want to allow a one-year lag for money to do its trick, was distinctly the most unstable in our recent history, marked by sharp fluctuations in output and wild gyrations of the rate of change of prices. As a result, the average deviation of the four-quarter changes in output was 3.3 percent, more than twice as large as in the period of less stable money growth. But the first stretch was also marked by well above average instability, with the contraction of 1954, the sharp recovery of 1955, and the new contraction in 1958, the sharpest in postwar history except for the present one. The variability of output is again 50 percent larger than in the middle period.

To be sure, in the recent episode serious exogenous shocks played a major role in the development of prices and possibly output, although the

same is not so readily apparent for the period 1953 to 1958. But, in any event, such extenuating circumstances are quite irrelevant to my point; for I am not suggesting that the stability of money was the major cause of economic instability—or at any rate, not yet! All I am arguing is that (i) there is no basis for the monetarists' suggestion that our postwar instability can be traced to monetary instability—our most unstable periods have coincided with periods of relative monetary stability; and (ii) stability of the money supply is not enough to give us a stable economy, precisely because there are exogenous disturbances.

Finally, let me mention that I have actually made an attempt at rerunning history to see whether a stable money supply would stabilize the economy, though in a way that I readily acknowledge is much inferior to the real thing, namely through a simulation with the *MPS*. The experiment, carried out in cooperation with Papademos, covered the relatively quiet period from the beginning of 1959 to the introduction of price-wage controls in the middle of 1971. If one eliminates all major sources of shocks, for example, by smoothing federal government expenditures, we found, as did Otto Eckstein in an earlier experiment, that a stable money growth of 3 percent per year does stabilize the economy, as expected. But when we allowed for all the historical shocks, the result was that with a constant money growth the economy was far from stable—in fact, it was distinctly less stable than actual experience, by a factor of 50 percent.

B. *The Overall Effectiveness of Postwar Stabilization Policies*

But even granted that a smooth money supply will not produce a very stable world and that there is therefore room for stabilization policies, monetarists will still argue that we should nonetheless eschew such policies. They claim, first, that allowing for unpredictably variable lags and unforseeable future shocks, we do not know enough to successfully design stabilization policies, and second, that the government would surely be incapable of choosing the appropriate policies or be politically willing to provide timely enforcement. Thus, in practice, stabilization policies will result in destabilizing the economy much of the time.

This view is supported by two arguments, one logical and one empirical. The logical argument is the one developed in Friedman's Presidential Address (1968). An attempt at stabilizing the economy at full employment is bound to be destabilizing because the full employment or natural rate is not known with certainty and is subject to shifts in time; and if we aim for the incorrect rate, the result must perforce be explosive inflation or deflation. By contrast, with a constant money supply policy, the economy will automatically hunt for, and eventually discover, that shifty natural rate, wherever it may be hiding.

This argument, I submit, is nothing but a debating ploy. It rests on the preposterous assumption that the only alternative to a constant money growth is the pursuit of a very precise unemployment target which will be adhered to indefinitely no matter what, and that if the target is off in the second decimal place, galloping inflation is around the corner. In reality, all that is necessary to pursue stabilization policies is a rough target range that includes the warranted rate, itself a range and not a razor edge; and, of course, responsible supporters of stabilization policies have long been aware of the fact that the target range needs to be adjusted in time on the basis of forseeable shifts in the warranted range, as well as in the light of emerging evidence that the current target is not consistent with price stability. It is precisely for this reason that I, as well as many other nonmonetarists, would side with monetarists in strenuous opposition to recent proposals for a target unemployment rate rigidly fixed by statute (although there is nothing wrong with Congress committing itself and the country to work toward the eventual achievement of some target unemployment rate through *structural* changes rather than aggregate demand policies).

Clearly, even the continuous updating of targets cannot guarantee that errors can be

avoided altogether or even that they will be promptly recognized; and while errors persist, they will result in some inflationary (or deflationary) pressures. But the growing inflation to which Friedman refers is, to repeat, a crawl not a gallop. One may usefully recall in this connection the experience of 1965–70 referred to earlier, with the further remark that the existence of excess employment was quite generally recognized at the time, and failure to eliminate it resulted overwhelmingly from political considerations and not from a wrong diagnosis.[3]

There remains then only the empirical issue: have stabilization policies worked in the past and will they work in the future? Monetarists think the answer is negative and suggest, as we have seen, that misguided attempts at stabilization, especially through monetary policies, are responsible for much of the observed instability. The main piece of evidence in support of this contention is the Great Depression, an episode well documented through the painstaking work of Friedman and Anna Schwartz, although still the object of dispute (see, for example, Peter Temin). But in any event, that episode while it may attest to the power of money, is irrelevant for present purposes since the contraction of the money supply was certainly not part of a comprehensive stabilization program in the post-Keynesian sense.

When we come to the relevant postwar period, the problem of establishing the success or failure of stabilization policies is an extremely taxing one. Many attempts have been made at developing precise objective tests, but in my view, none of these is of much value, even though I am guilty of having contributed to them in one of my worst papers (1964). Even the most ingenious test, that suggested by Victor Argy, and relying on a comparison of the variability of income with that of the velocity of circulation, turns out to be valid only under highly unrealistic restrictive assumptions.

Dennis Starleaf and Richard Floyd have proposed testing the effectiveness of stabilization by comparing the stability of money growth with that of income growth, much as I have done above for the United States, except that they apply their test to a cross section of industrialized countries. They found that for a sample of 13 countries, the association was distinctly positive. But this test is again of little value. For while a negative association for a given country, such as suggested by my *U.S.* test, does provide some weak indication that monetary activism helped rather than hindered, the finding of a positive association across countries proves absolutely nothing. It can be readily shown, in fact, that, to the extent that differential variability of income reflects differences in the character of the shocks—a most likely circumstance for their sample—successful stabilization also implies a positive correlation between the variability of income and that of money.

But though the search for unambiguous quantitative tests has so far yielded a meager crop, there exists a different kind of evidence in favor of Keynesian stabilization policies which is impressive, even if hard to quantify. To quote one of the founding fathers of business cycle analysis, Arthur Burns, writing in 1959, "Since 1937 we have had five recessions, the longest of which lasted only 13 months. There is no parallel for such a sequence of mild—or such a sequence of brief—contractions, at least during the past hundred years in our country" (p. 2). By now we can add to that list the recessions of 1961 and 1970.

There is, furthermore, evidence that very similar conclusions hold for other industrialized countries which have made use of stabilization policies; at any rate that was the prevailing view among participants to an international conference held in 1967 on the subject, "Is the busi-

[3]Friedman's logical argument against stabilization policies and in favor of a constant money growth rule is, I submit, much like arguing to a man from St Paul wishing to go to New Orleans on important business that he would be a fool to drive and should instead get himself a tub and drift down the Mississippi: that way he can be pretty sure that the current will eventually get him to his destination, whereas, if he drives, he might make a wrong turn and, before he notices he will be going further and further away from his destination and pretty soon he may end up in Alaska, where he will surely catch pneumonia and he may never get to New Orleans!

ness cycle obsolete?" (see Martin Bronfenbrenner, editor). No one seemed to question the greater postwar stability of all Western economies—nor is this surprising when one recalls that around that time business cycle specialists felt so threatened by the new-found stability that they were arguing for redefining business cycles as fluctuations in the *rate of growth* rather than in the *level* of output.

It was recognized that the reduced severity of fluctuations might in part reflect structural changes in the economy and the effect of stronger built-in stabilizers, inspired, of course, by the Keynesian analysis. Furthermore, the greater stability in the United States, and in other industrialized countries, are obviously not independent events. Still, at least as of the time of that conference, there seemed to be little question and some evidence that part of the credit for the greater stability should go to the conscious and on balance, successful endeavor at stabilizing the economy.

V. The Case of Supply Shocks and the 1974–76 Episode

A. *Was the 1974 Depression Due to Errors of Commission or Omission?*

In pointing out our relative postwar stability and the qualified success of stabilization policies, I have carefully defined the postwar period as ending somewhere in 1973. What has happened since that has so tarnished the reputation of economists? In facing this problem, the first question that needs to be raised is whether the recent combination of unprecedented rates of inflation as well as unemployment must be traced to crimes of commission or omission. Did our monetary and fiscal stabilization policies misfire, or did we instead fail to use them?

We may begin by establishing one point that has been blurred by monetarists' blanket indictments of recent monetary policy: the virulent explosion that raised the four-quarter rate of inflation from about 4 percent in 1972 to 6.5 percent by the third quarter of 1973, to 11.5 percent in 1974 with a peak quarterly rate of 13.5, can in no way be traced to an excessive, or to a disorderly, growth of the money supply. As already mentioned, the average rate of money growth from the beginning of 1970 to the second half of 1974 was close to 7 percent. To be sure, this was a high rate and could be expected sooner or later to generate an undesirably high inflation —but how high? Under any reasonable assumption one cannot arrive at a figure much above 6 percent. This might explain what happened up to the fall of 1973, but not from the third quarter of 1973 to the end of 1974, which is the really troublesome period. Similarly, as was indicated above, the growth of money was reasonably smooth over this period, smoother than at any other time in the postwar period, staying within a 2 percent band. Hence, the debacle of 1974 can just not be traced to an erratic behavior of money resulting from a misguided attempt at stabilization.

Should one then conclude that the catastrophe resulted from too slavish an adherence to a stable growth rate, forsaking the opportunity to use monetary policy to stabilize the economy? In one sense, the answer to this question must in my view be in the affirmative. There is ample ground for holding that the rapid contraction that set in toward the end of 1974, on the heels of a slow decline in the previous three quarters, and which drove unemployment to its 9 percent peak, was largely the result of the astronomic rise in interest rates around the middle of the year. That rise in turn was the unavoidable result of the Fed's stubborn refusal to accommodate, to an adequate extent, the exogenous inflationary shock due to oil, by letting the money supply growth exceed the 6 percent rate announced at the beginning of the year. And this despite repeated warnings about that unavoidable result (see, for example, the author 1974).

Monetarists have suggested that the sharp recession was not the result of too slow a monetary growth throughout the year, but instead of the deceleration that took place in the last half of 1974, and early 1975. But this explanation just does not stand up to the facts. The fall in the quarterly growth of money in the third and fourth quarters was puny, especially on the basis of

revised figures now available: from 5.7 percent in the second to 4.3 and 4.1—hardly much larger than the error of estimate for quarterly rates! To be sure, in the first quarter of 1975 the growth fell to .6 percent. But, by then, the violent contraction was well on its way—between September 1974 and February 1975, industrial production fell at an annual rate of 25 percent. Furthermore, by the next quarter, monetary growth had resumed heftily. There is thus no way the monetarist proposition can square with these facts unless their long and variable lags are so variable that they sometimes turn into substantial leads. But even then, by anybody's model, a one-quarter dip in the growth of money could not have had a perceptible effect.

B. *What Macro Stabilization Policies Can Accomplish, and How*

But recognizing that the adherence to a stable money growth path through much of 1974 bears a major responsibility for the sharp contraction does not per se establish that the policy was mistaken. The reason is that the shock that hit the system in 1973–74 was not the usual type of demand shock which we have gradually learned to cope with, more or less adequately. It was, instead, a supply or price shock, coming from a cumulation of causes, largely external. This poses an altogether different stabilization problem. In particular, in the case of demand shocks, there exists in principle an ideal policy which avoids all social costs, namely to offset completely the shock thus, at the same time, stabilizing employment and the price level. There may be disagreement as to whether this target can be achieved and how, but not about the target itself.

But in the case of supply shocks, there is no miracle cure—there is no macro policy which can both maintain a stable price level and keep employment at its natural rate. To maintain stable prices in the face of the exogenous price shock, say a rise in import prices, would require a fall in all domestic output prices; but we know of no macro policy by which domestic prices can be made to fall except by creating enough slack, thus putting downward pressure on wages. And the amount of slack would have to be substantial in view of the sluggishness of wages in the face of unemployment. If we do not offset the exogenous shock completely, then the initial burst, even if activated by an entirely transient rise in some prices, such as a once and for all deterioration in the terms of trade, will give rise to further increases, as nominal wages rise in a vain attempt at preserving real wages; this secondary reaction too can only be cut short by creating slack. In short, once a price shock hits, there is no way of returning to the initial equilibrium except after a painful period of both above equilibrium unemployment and inflation.

There are, of course, in principle, policies other than aggregate demand management to which we might turn, and which are enticing in view of the unpleasant alternatives offered by demand management. But so far such policies, at least those of the wage-price control variety, have proved disappointing. The design of better alternatives is probably the greatest challenge presently confronting those interested in stabilization. However, these policies fall outside my present concern. Within the realm of aggregate demand management, the only choice open to society is the cruel one between alternative feasible paths of inflation and associated paths of unemployment, and the best the macroeconomist can offer is policies designed to approximate the chosen path.

In light of the above, we may ask: is it conceivable that a constant rate of growth of the money supply will provide a satisfactory response to price shocks in the sense of giving rise to an unemployment-inflation path to which the country would object least?

C. *The Monetarist Prescription: Or, Constant Money Growth Once More*

The monetarists are inclined to answer this question affirmatively, if not in terms of the country's preferences, at least in terms of the preferences they think it should have. This is evidenced by their staunch support of a continuation of the 6 percent or so rate of growth through

1974, 1975, and 1976.

Their reasoning seems to go along the following lines. The natural rate hypothesis implies that the rate of inflation can change only when employment deviates from the natural rate. Now suppose we start from the natural rate and some corresponding steady rate of inflation, which without loss of generality can be assumed as zero. Let there be an exogenous shock which initially lifts the rate of inflation, say, to 10 percent. If the Central Bank, by accommodating this price rise, keeps employment at the natural rate, the new rate of 10 percent will also be maintained and will in fact continue forever, as long as the money supply accommodates it. The only way to eliminate inflation is to increase unemployment enough, above the natural rate and for a long enough time, so that the cumulated reduction of inflation takes us back to zero. There will of course be many possible unemployment paths that will accomplish this. So the next question is: Which is the least undesirable?

The monetarist answer seems to be—and here I confess that attribution becomes difficult—that it does not make much difference because, to a first approximation, the cumulated amount of unemployment needed to unwind inflation is independent of the path. If we take more unemployment early, we need to take less later, and conversely. But then it follows immediately that the specific path of unemployment that would be generated by a constant money growth is, if not better, at least as good as any other. Corollary: a constant growth of money is a satisfactory answer to supply shocks just as it is to demand shocks—as well as, one may suspect, to any other conceivable illness, indisposition, or disorder.

D. *Why Constant Money Growth Cannot Be the Answer*

This reasoning is admirably simple and elegant, but it suffers from several flaws. The first one is a confusion between the price level and its rate of change. With an unchanged constant growth of the nominal money stock, the system will settle back into equilibrium not when the rate of inflation is back to zero but only when, in addition, the price level itself is back to its initial level. This means that when inflation has finally returned back to the desired original rate, unemployment cannot also be back to the original level but will instead remain above it as long as is necessary to generate enough deflation to offset the earlier cumulated inflation. I doubt that this solution would find many supporters and for a good reason; it amounts to requiring that none of the burden of the price shock should fall on the holder of long-term money fixed contracts—such as debts—and that all other sectors of society should shoulder entirely whatever cost is necessary to insure this result. But if, as seems to be fairly universally agreed, the social target is instead to return the system to the original rate of inflation—zero in our example—then the growth of the money supply cannot be kept constant. Between the time the shock hits and the time inflation has returned to the long-run level, there must be an additional increase in money supply by as much as the price level or by the cumulant of inflation over the path.

A second problem with the monetarists' argument is that it implies a rather special preference function that depends only on cumulated unemployment. And, last but not least, it requires the heroic assumption that the Phillips curve be not only vertical in the long run but also linear in the short run, an assumption that does not seem consistent with empirically estimated curves. Dropping this last assumption has the effect that, for any given social preference, there will be in general a unique optimal path. Clearly, for this path to be precisely that generated by a constant money growth, would require a miracle—or some sleight of the invisible hand!

Actually, there are grounds for holding that the unemployment path generated by a constant money growth, even if temporarily raised to take care of the first flaw, could not possibly be close to an optimal. This conclusion is based on an analysis of optimal paths, relying on the type of linear welfare function that appears to underlie the monetarists' argument, and which is also a straightforward generalization of Okun's fa-

mous "economic discomfort index." That index (which according to Michael Lovell appears to have some empirical support) is the sum of unemployment and inflation. The index used in my analysis is a weighted average of the cumulated unemployment and cumulated inflation over the path. The weights express the relative social concern for inflation versus unemployment.

Using this index, it has been shown in a forthcoming thesis of Papademos that, in general, the optimum policy calls for raising unemployment at once to a certain critical level and keeping it there until inflation has substantially abated. The critical level depends on the nature of the Phillips curve and the relative weights, but does not depend significantly on the initial shock—as long as it is appreciable. To provide an idea of the order of magnitudes involved, if one relies on the estimate of the Phillips curve reported in my joint paper with Papademos (1975), which is fairly close to vertical and uses Okun's weights, one finds that (i) at the present time, the noninflationary rate of unemployment corresponding to a 2 percent rate of inflation can be estimated at 5.6 percent, and (ii) the optimal response to a large exogenous price shock consists in increasing unemployment from 5.6 to only about 7 percent. That level is to be maintained until inflation falls somewhat below 4 percent; it should then be reduced slowly until inflation gets to 2.5 (which is estimated to take a couple of years), and rapidly thereafter. If, on the other hand, society were to rate inflation twice as costly as unemployment, the initial unemployment rate becomes just over 8 percent, though the path to final equilibrium is then shorter. These results seem intuitively sensible and quantitatively reasonable, providing further justification for the assumed welfare function, with its appealing property of summarizing preferences into a single readily understandable number.

One important implication of the nature of the optimum path described above is that a constant money growth could not possibly be optimal while inflation is being squeezed out of the system, regardless of the relative weights attached to unemployment and inflation. It would tend to be prevailingly too small for some initial period and too large thereafter.

One must thus conclude that the case for a constant money growth is no more tenable in the case of supply shocks than it is in the case of demand shocks.

VI. Conclusion

To summarize, the monetarists have made a valid and most valuable contribution in establishing that our economy is far less unstable than the early Keynesians pictured it and in rehabilitating the role of money as a determinant of aggregate demand. They are wrong, however, in going as far as asserting that the economy is sufficiently shockproof that stabilization policies are not needed. They have also made an important contribution in pointing out that such policies might in fact prove destabilizing. This criticism has had a salutary effect on reassessing what stabilization policies can and should do, and on trimming down fine-tuning ambitions. But their contention that postwar fluctuations resulted from an unstable money growth or that stabilization policies decreased rather than increased stability just does not stand up to an impartial examination of the postwar record of the United States and other industrialized countries. Up to 1974, these policies have helped to keep the economy reasonable stable by historical standards, even though one can certainly point to some occasional failures.

The serious deterioration in economic stability since 1973 must be attributed in the first place to the novel nature of the shocks that hit us, namely, supply shocks. Even the best possible aggregate demand management cannot offset such shocks without a lot of unemployment together with a lot of inflation. But, in addition, demand management was far from the best. This failure must be attributed in good measure to the fact that we had little experience or even an adequate conceptual framework to deal with such shocks; but at least from my reading of the record, it was also the result of failure to use stabilization policies, including too slavish adherence to the monetarists' constant money

growth presciption.

We must, therefore, categorically reject the monetarist appeal to turn back the clock forty years by discarding the basic message of *The General Theory*. We should instead concentrate our efforts in an endeavor to make stabilization policies even more effective in the future than they have been in the past

REFERENCES

L. C. Andersen and K. M. Carlson, "A Monetarist Model for Economic Stabilization," *Fed. Reserve Bank St Louis Rev*, Apr 1970, *52*, 7–25.

——— **and J. L. Jordan,** "Monetary and Fiscal Action: A Test of Their Relative Importance in Economic Stabilization," *Fed. Reserve Bank St. Louis Rev.*, Nov 1968, *50*, 11–23.

V. Argy, "Rules, Discretion in Monetary Management, and Short-Term Stability," *J Money, Credit, Banking*, Feb. 1971, *3*, 102–22.

W. J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quart. J. Econ.*, Nov. 1952, *66*, 545–56.

R. G. Bodkin, "Real Wages and Cyclical Variations in Employment: A Reexamination of the Evidence," *Can. J. Econ.*, Aug. 1969, *2*, 353–74.

Martin Bronfenbrenner, *Is the Business Cycle Obsolete?*, New York 1969.

A. F. Burns, "Progress Towards Economic Stability," *Amer. Econ Rev.*, Mar. 1960, *50*, 1–19.

J. T. Dunlop, "The Movement of Real and Money Wage Rates," *Econ. J.*, Sept. 1938, *48*, 413–34.

O. Eckstein and R. Brinner, "The Inflation Process in the United States," in Otto Eckstein, ed., *Parameters and Policies in the U.S. Economy*, Amsterdam 1976.

R. C. Fair, "On Controlling the Economy to Win Elections," unpub. paper, Cowles Foundation 1975.

M. S. Feldstein, "Temporary Layoffs in the Theory of Unemployment," *J. Polit. Econ.*, Oct. 1976, *84*, 937–57.

S. Fischer, "Long-term Contracts, Rational Expectations and the Optimal Money Supply Rule," *J. Polit. Econ.*, forthcoming.

B. M. Friedman, "Rational Expectations Are Really Adaptive After All," unpub. paper, Harvard Univ. 1975.

Milton Friedman, *A Theory of the Consumption Function*, Princeton 1957.

———, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

———, "The Demand for Money: Some Theoretical and Empirical Results," in his *The Optimum Quantity of Money, and Other Essays*, Chicago 1969.

——— **and A. Schwartz,** *A Monetary History of the United States 1867–1960*, Princeton 1963.

S. Goldfeld, "The Demand for Money Revisited," *Brookings Papers*, Washington 1973, *3*, 577–646.

R. J. Gordon, "Recent Developments in the Theory of Inflation and Unemployment," *J Monet. Econ.*, Apr. 1976, *2*, 185–219.

J. R. Hicks, "Mr. Keynes and the "Classics"; A Suggested Interpretation," *Econometrica*, Apr. 1937, *5*, 147–59.

John Maynard Keynes, *The General Theory of Employment, Interest and Money*, New York 1935

R. G. Lipsey, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A Further Analysis," *Economica*, Feb. 1960, *27*, 1–31.

M. Lovell, "Why Was the Consumer Feeling So Sad?," *Brookings Papers*, Washington 1975, 2, 473–79.

R. E. Lucas, Jr., "Econometric Policy Evaluation: A Critique," *J. Monet. Econ.*, suppl. series, 1976, *1*, 19–46.

———, "Expectations and the Neutrality of Money," *J. Econ Theory*, Apr. 1972, *4*, 103–24.

M. Miller and D. Orr, "A Model of the Demand for Money by Firms," *Quart. J. Econ.*, Aug. 1966, *80*, 413–35.

F. Modigliani, "Liquidity Preference and the Theory of Interest and Money," *Econo-*

metrica, Jan. 1944, *12*, 45–88.

———, "New Development on the Oligopoly Front," *J. Polit. Econ.*, June 1958, *66*, 215–33.

———, "The Monetary Mechanism and Its Interaction with Real Phenomena," *Rev. Econ. Statist.*, Feb. 1963, *45*, 79–107.

———, " Some Empirical Tests of Monetary Management and of Rules versus Discretion," *J. Polit. Econ.*, June 1964, *72*, 211–45.

———, "The 1974 Report of the President's Council of Economic Advisers: A Critique of Past and Prospective Policies," *Amer. Econ. Rev.*, Sept. 1974, *64*, 544–77.

———, "The Life Cycle Hypothesis of Saving Twenty Years Later," in Michael Parkin, ed., *Contemporary Issues in Economics*, Manchester 1975.

——— **and A. Ando,** "The Relative Stability of Monetary Velocity and the Investment Multiplier," *Amer. Econ. Rev.*, Sept. 1965, *55*, 693–728.

——— **and** ———, "Impacts of Fiscal Actions on Aggregate Income and the Monetarist Controversy: Theory and Evidence," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976.

——— **and R. Brumberg,** "Utility Analysis and the Consumption Function: Interpretation of Cross-Section Data," in Kenneth Kurihara, ed., *Post-Keynesian Economics*, New Brunswick 1954.

——— **and L. Papademos,** "Targets for Monetary Policy in the Coming Years," *Brookings Papers*, Washington 1975, *1*, 141–65.

——— **and** ———, "Monetary Policy for the Coming Quarters: The Conflicting Views," *New Eng. Econ. Rev.*, Mar./Apr. 1976, 2–35.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

W. D. Nordhaus, "The Political Business Cycle," *Rev. Econ. Stud.*, Apr. 1975, *42*, 169–90.

A. M. Okun, "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers*, Washington 1975, *2*, 351–90.

D. O'Neill, "Directly Estimated Multipliers of Monetary and Fiscal Policy," doctoral thesis in progress, M.I.T.

L. Papademos, "Optimal Aggregate Employment Policy and Other Essays," doctoral thesis in progress, M.I.T.

Edmond S. Phelps, "Money-Wage Dynamics and Labor-Market Equilibrium," *J. Polit. Econ.*, July/Aug. 1968, *76*, 678–711.

——— **et al.,** *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

A. W. Phillips, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, Nov. 1958, *25*, 283–99.

T. J. Sargent, "A Classical Macroeconomic Model for the United States," *J. Polit. Econ.*, Apr. 1976, *84*, 207–37.

——— **and N. Wallace,** "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–57

D. Starleaf and R. Floyd, "Some Evidence with Respect to the Efficiency of Friedman's Monetary Policy Proposals," *J. Money, Credit, Banking*, Aug. 1972, *4*, 713–22.

Peter Temin, *Did Monetary Forces Cause the Great Depression?*, New York 1976.

James Tobin, *Essays in Economics: Vol. 1, Macroeconomics*, Chicago 1971

Should Government Subsidize Risky Private Projects?

By Joram Mayshar*

With the growth of public involvement in economic life, the subsidization of private projects has become a common practice. In many countries, government subsidization of new private ventures through easy loans, matching grants, tax rebates, protective tariffs, and the like has become of far greater significance for industrial and agricultural growth than direct public investment. In this paper I shall examine whether considerations of risk provide a justification for subsidies to private projects.

The theoretical literature dealing with this topic is very limited. Indeed only cursory remarks in the debate concerning the social rate of discount have been devoted to the practice of subsidization. The relevant issues in the debate on the social rate of discount can be represented by the viewpoints of Jack Hirshleifer on the one hand, and Kenneth Arrow and Robert Lind on the other.[1] Hirshleifer's general conclusion is that if capital markets are perfect, public projects should be evaluated using the risk-adjusted discount rate employed in the private sector for projects with a "comparable" type of risk Arrow and Lind accept Hirshleifer's proposition, but argue that it is generally inapplicable since capital markets are not perfect. Their own principal conclusion is that if a public project is independent of all other projects in the economy, then, whether or not the capital markets are perfect, the government should use a riskless rate of discount. Their underlying reasoning (which is quite separate from the often cited risk-pooling argument) is that the total cost of risk bearing of an independent public project is made negligible by the spreading of the risk over the entire population of taxpayers.

While on the issue of the social discount rate, Hirshleifer and Arrow-Lind reach different conclusions due to different underlying assumptions; on the issue of subsidization they differ without an apparent difference in assumptions. Their conflicting conclusions on subsidization can best be illustrated by the example suggested by Arrow and Lind which compares the social desirability of two alternative investment schemes, both employing the same resources in an independent project. If the project were to be undertaken by a private firm, its expected rate of (net) return would be 10 percent but the risk-adjusted rate would be only 5 percent. On the other hand, if government undertook the project, its expected rate of return would be 6 percent which is also the market's riskless rate of return. As Arrow and Lind make clear, the fact that the private firm adjusts for risk even though the project is independent indicates that the capital markets are imperfect and do not facilitate the sufficient spreading of the private project's risk. Hirshleifer, considering a similar example, parenthetically commented that the adoption of the public project can be justified only as a "second best"—"For, granted the premises it would clearly be most efficient for the government to borrow in order to subsidize the higher-expected-yield private investment . . . rather than [to undertake the] lower-yield public investment" (p. 270).

Arrow and Lind have taken issue with this recommendation for subsidization. They argue that because of its independence, the costs of risk bearing of the public project can be ignored whereas the risk-bearing costs of the private project are real costs to the economy. By

*Lecturer in economics, Hebrew University, Jerusalem. I am very much indebted to Peter Diamond for his help and encouragement

[1]A critical survey of the debate on the social rate of discount under uncertainty can be found in the article by Martin Bailey and Michael Jensen

the Kaldor-Hicks principle of potential compensation, the public project is thus to be preferred. They reject the recommendation for subsidization on the grounds that: ". . . a program of providing direct subsidies to encourage more private investment does not alter the costs of risk-bearing and, therefore, will encourage investments which are inefficient when the costs of risk are considered" (p. 375).

The major point of this paper is that subsidization of private risky projects *is* a justified, even necessary, practice given two conditions: that the capital markets are incomplete and that income taxation exists. Indeed, if to the above example we introduce income taxation at a rate of 50 percent, then we find that it is the private project which ought to be undertaken in preference to the public project. In order that the private project be undertaken it must be subsidized by the government. By way of a guideline we find below that the required subsidy could take the form of a loan at a reduced rate of 3.5 percent which would be low enough to encourage further expansion of the private project.

The intuitive reason for our general conclusion in favor of the subsidization of private projects rests on an observation made by Evsey Domar and Richard Musgrave: "By imposing an income tax on the investor, the treasury appoints itself as his partner" (p. 389). Therefore when taxes exist and when, due to capital market deficiencies, the government's attitude towards risk differs from that of the private investor, it is in the government's interest that the investment decision reflect its own concerns in the assumed partnership. In particular, if the government is more neutral toward risk than the private investor, then it ought to subsidize the project in order to further its level of investment.

In the controversy between Hirshleifer favoring subsidies, and Arrow and Lind arguing against subsidies, we thus end somewhere in the middle—favoring subsidies but for different reasons than those suggested by Hirshleifer. It should be emphasized however that my argument does not differ in its general approach from that of Arrow and Lind. Rather, since a new consideration of taxation is introduced it should be regarded as complementary to theirs. However, whereas Arrow and Lind argue that "The program which produces the desired result is one to insure private investments [rather than to subsidize them]" (p. 375), we find here that it is income taxation, itself a form of insurance, which creates the very case for subsidization.

I. Sandmo's "Farm Model"

The preceding heuristic argument in favor of subsidization will be analyzed more rigorously by use of a very stylized model. The model is based on Agnar Sandmo's ingenuous representation of an "unincorporated economy," itself an adaptation of Peter Diamond's contrasting "stock-market economy." The model represents a polar case of an economy without a system of capital markets; its results, however, are clearly not limited to that polar case only. We can think of Sandmo's model as representing a farm economy in which due to the lack of capital markets, the individual farmers have to bear the entire risks of production. The only financial market in the farm economy is a loan market run by the central bank. The loan market is assumed to operate costlessly and to involve no risks of default. We further assume that only a single good exists in that economy ("corn") and that consumption takes place in a single future period.[2] The source of the uncertainty with regard to the future state of nature is "technological," i.e., it is like the weather, independent of actions in the economy.[3]

Let there be n individual farmers (or producer-consumers), each one operating a farm (henceforth, a firm) with a given technology such that for current inputs of y_j, the future's uncertain output is $f_j(y_j)\phi_j(\theta)$, where

[2]In Sandmo's original model, present period consumption is included. I exclude this extension so as not to confound the main argument with the possible (yet not necessary) intertemporal distortions caused by taxation.

[3]The assumption of technological uncertainty is significant in our analysis. We neglect here entirely the often made argument which justifies subsidization on the grounds that many of the "nontechnological" risks (such as the risk of temporary insolvency) facing private inventors are not "social" risks.

θ denotes the state of nature.[4] The term $\phi_j(\theta)$ represents a pattern of returns across the various states of nature and can be considered as a "composite good."[5] This composite good consists of $\phi_j(\theta')$ units of the θ' contingent commodity for each state of nature θ'. (To recall, the Arrow-Debreu θ' contingent commodity offers the delivery of a unit of future consumption if and only if the state θ' occurs.) We can thus think of firm j as producing $f_j(y_j)$ units of the composite good $\phi_j(\theta)$. The "production function" $f_j(y_j)$ is assumed to satisfy the standard assumptions (i.e., $f_j'(y_j) > 0$, $f_j''(y_j) \le 0$ for all $y_j \ge 0$).

In addition to investment in production, each farmer can make or receive loans from the central bank at a given rate of safe return r. His budget constraint is thus:

$$(1) \qquad y_j + b_j = w_j \qquad j = 1, \ldots, n$$

where b_j is the amount of bonds held (i.e., of loans made out), and w_j the initial wealth of farmer j. The farmer's future gross income will then be $f_j(y_j)\phi_j(\theta) + [1 + r]b_j$. We assume now that a uniform income tax of a rate t ($0 \le t < 1$) exists in the economy. As the basis for income taxation we take a measure of net income —the excess of gross future income over initial wealth.[6] We furthermore assume that all the government's revenue $G(\theta)$, derived from tax collection, from public production, and from profits of the central bank, is redistributed to the farmers according to preassigned fractions τ_j, $j = 1, \ldots, n$ $(\Sigma_{j=1}^{n} \tau_j = 1)$. The individual farmer's future consumption is thus:

$$(2) \qquad c_j(\theta) = [1 - t][f_j(y_j)\phi_j(\theta) + [1 + r]b_j] + tw_j + \tau_j G(\theta) \quad j = 1, \ldots, n$$

Each farmer j is assumed to attempt to maximize the expected utility of his future consumption $E_j u_j(c_j(\theta))$ using his own subjective probability assessments. The utility function $u_j(c_j)$ is assumed to satisfy the standard assumptions (i.e., $u'_j(c_j) > 0$, $u''_j(c_j) \le 0$ for any $c_j \ge 0$).[7]

We denote:

$$(3) \qquad S_j^i = \frac{E_i u_i'(c_i(\theta))\phi_j(\theta)}{E_i u_i'(c_i(\theta))}$$

The function S_j^i will be termed here as the "subjective valuation" by individual i of a marginal unit of the composite good $\phi_j(\theta)$. This function which is but the marginal rate of substitution between the composite good $\phi_j(\theta)$ and safe future consumption can be alternatively interpreted as the subjective marginal-certainty equivalent of the composite $\phi_j(\theta)$.[8]

The first-order conditions for an individual j maximizing his expected utility, constrained by (1) and (2) and taking $\tau_j G(\theta)$ as given, can now be written as:

$$(4) \qquad f'_j(y_j)S_j = 1 + r \quad , j = 1, \ldots, n$$

The condition (4) is assumed to define implicitly the investment level which will be adopted by farmer-investor j. This condition has a standard interpretation requiring the equation of the (subjective) value of the marginal product of the investment resource to its price (in terms of the alternative future consumption foregone).[9]

Following Sandmo, public production is introduced to the farm model in a stylized way by assuming that the government can produce in the

[4]The multiplicative form of the production function is *not* essential for our argument. We could, for example, allow firm j to produce $\Sigma_{k=1}^{K} f_{jk}\phi_{jk}(\theta)$ future output for $h_j(f_{j1}, \ldots, f_{jK})$ present inputs. When K equals the number of states of nature such a formulation will overburden the exposition without significantly changing our results.

[5]If the probability assessments of all individuals were the same (which, unlike Sandmo, we do not assume), a natural normalization of the composite goods would be to set $E\phi_j(\theta) = 1$ for all j.

[6]The results of the paper are insensitive to the particular form of taxation assumed here. However it is significant that the tax be imposed at the same rate on income from production and from loans, so that no distortions are introduced.

[7]We will ignore throughout the paper the possibility of default, i.e., of $c_j(\theta) < 0$.

[8]If the certainty equivalent π_{ij} is defined by $E_i u_i(c_i(\theta) + s\phi_j(\theta)) \equiv E_i u_i(c_i(\theta) + \pi_{ij}(s))$, then it is simple to show that $S_j^i = \pi'_{ij}(0)$.

[9]Alternatively, one can use risk-adjusted rates of return to express (4) as $f'_j(y_j) = [1 + r][S_j^j]^{-1}$. Sandmo uses that approach and interprets $[S_j^j]^{-1}$ as a "risk margin."

same risk classes as the private sector. That is, public firm k for inputs z_k would yield a future output of $g_k(z_k)\phi_k(\theta)$, the functions $g_k(z_k)$ satisfying the standard assumptions, $k = 1, \ldots, n$. Public investment is assumed to be financed by debt, through the central bank. The equilibrium riskless rate of return will be determined then so as to just exhaust the available investment resources; i.e.,

$$\sum_{j=1}^{n} y_j + \sum_{j=1}^{n} z_j = \sum_{j=1}^{n} w_j \tag{5}$$

II. The Social Optimum

Following Diamond's concept of a "constrained Pareto optimum," we define the "social optimum" as the allocation of resources that a social planner maximizing social welfare would select, given the existing market structure. The planner has thus to choose optimal levels for private and public investments (y_j, z_j for $j = 1, \ldots, n$) and to select the return r so as to maximize social welfare, assumed here to be given by the individualistic function $\Sigma^n_{j=1} \lambda_j E_j u_j(c_j(\theta))$. For technical reasons of easing the exposition, we furthermore assume that the government can redistribute initial resources in a lump sum fashion (or alternatively that initial wealth is already optimally distributed). The constraints on the social planner are the resource constraint (5), the definition of individuals' consumption (2), and the definition of the government's combined revenues:

$$G(\theta) = \sum_{j=1}^{n} g_j(z_j)\phi_j(\theta) + t \sum_{j=1}^{n} f_j(y_j)\phi_j(\theta) - [1 + r][1 - t] \sum_{j=1}^{n} b_j - t \sum_{j=1}^{n} w_j \tag{6}$$

The assumption of the possibility of lump sum initial wealth redistribution yields the first-order condition:

$$\lambda_j E_j u'_j(c_j(\theta)) = \nu \qquad j = 1, \ldots, n \tag{7}$$

With equation (7), the other first-order conditions in addition to (5) become:

$$f'_j(y_j)\Big[[1 - t]S^j_j + t \sum_{i=1}^{n} \tau_i S^i_j\Big] = 1 + \rho \qquad j = 1, \ldots, n \tag{8}$$

$$g'_j(z_j) \sum_{i=1}^{n} \tau_i S^i_j = 1 + \rho \qquad j = 1, \ldots, n \tag{9}$$

We assume that these first-order conditions do indeed define the social optimum.

The shadow price $1 + \rho$ in conditions (8) and (9) measures the social value of present resources in the economy in terms of safe future consumption. Thus ρ can be considered as the social rate of discount. It is interesting to note here that if foreign borrowing and lending at a uniform rate r^* were open to the government, then the social discount rate ρ would be replaced in the conditions (8) and (9) by the international rate r^*.[10]

Conditions (8) and (9) provide the socially optimal rules for private and public investment. We now briefly examine the public investment rule (9) and show that it can be considered as a generalization of both the investment criteria proposed by Hirshleifer and by Arrow and Lind.

In analogy to the interpretation of S^j_j for the individual producer, the term $\Sigma^n_{i=1} \tau_i S^i_j$ can be interpreted as the "social valuation" of the composite good $\phi_j(\theta)$. Condition (9) will then require the equalization of the social value of marginal investment in each public firm to the social price of investment resources—the gross social rate of return. The social valuation of any composite good $\phi_j(\theta)$ is a weighted average of consumers' subjective valuations of that composite, the weights being their (marginal) shares in that composite good as received from the government.

If a stock market existed in which all individuals could costlessly trade shares in the stock of firm j, then we would obtain $S^i_j = S^j_j$ for all i. This result can be explained by noting that the market trading in the stock of firm j can be

[10] In this case the resource constant (5) would drop and the government's combined revenue in (6) would increase by $[1 + r^*]\left[\sum_{j=1}^{n} w_j - \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} z_j\right]$

regarded as tantamount to direct trading in the composite good $\phi_j(\theta)$. Such trading equates all individuals' subjective valuations of this composite good to its (implicit) market price. The social valuation of the composite $\phi_j(\theta)$ would then be equal to the private subjective valuation. Following Diamond and Sandmo we then obtain that if the market rate of return r were to equal the social rate of discount ρ, the optimal rule for investment by the public firm j would be to imitate the investment rule of the private firm in that same risk class. This is exactly the investment rule advanced by Hirshleifer.

On the other hand, rule (9) also implies the Arrow-Lind recommendation that in the case of independent projects the government should be neutral towards risk, provided the population is large enough.[11] When a public project j is statistically independent of all other projects in the economy (in particular there is then no private project j in the same risk class), and when only a small share in it accrues to each individual (assuming $\tau_i = 1/n$ for all i), the subjective valuations can be shown to be approximated by:[12]

$$(10) \quad S_j^i \simeq E_i\phi_j(\theta) - \frac{1}{n}A_i(c_i(\theta))g_j(z_j) \cdot [\mathrm{Var}_i\phi_j(\theta)] \qquad i = 1, \ldots, n, \quad (i \neq j)$$

where

$$(11) \qquad A_i(c_i(\theta)) = -\frac{E_iu_i''(c_i(\theta))}{E_iu_i'(c_i(\theta))}$$

is a measure of risk aversion comparable to the Arrow-Pratt measure of absolute risk aversion. If n is large enough and if all assessments $E_i\phi_j(\theta)$ are the same, then in computing the social valuation of $\phi_j(\theta)$, we would get $\Sigma_{i=1}^n \tau_iS_j^i = E\phi_j(\theta)$. Thus condition (9) for optimal public investment indeed reduces to the Arrow-Lind condition of risk neutrality in the case of independent projects.

For our purposes, however, the major interest lies not with the rule for public investment, but rather with the socially optimal rule for private investment given by (8).

III. The Case for Subsidization

The striking feature in the optimal rule for private investment (8) is the term $[[1-t]S_j^j + t\Sigma_{i=1}^n\tau_iS_j^i]$. That term represents the valuation of the composite good $\phi_j(\theta)$ which is required from a social point of view for optimal production by the private firm j. This valuation is a weighted average of the private and social valuations of the composite; the weights are the shares in the private output accorded to each of the respective "partners" by the tax structure. Condition (8) thus formalizes the argument given in the introduction that the government's attitude towards risk should be taken into account in the determination of the private level of investment.

Defining now the "excess social value" of $\phi_j(\theta)$ as

$$(12) \quad \eta_j = \frac{\sum_{i=1}^n \tau_iS_j^i}{S_j^j} - 1 \qquad j = 1, \ldots, n$$

Condition (8) for the optimal private investment can then be rewritten as

$$(13) \qquad f_j'(v_j)S_j^j[1 + t\eta_j] = 1 + \rho \qquad j = 1, \ldots, n$$

From condition (13) it is clear that if no taxation existed, provided that the market rate of return is properly set (i.e., $r = \rho$), the private investment rule (4) will be socially optimal. Thus the social optimum could be decentralized. In particular, no form of subsidization will then

[11] The fact that the Arrow-Lind argument is just a subcase of (9) was somehow overlooked by Sandmo (see p. 300).

[12] From the definition of S_j^i in (3), to verify (10) and (11) we have to show that $E_iu_i'(c_i(\theta))\psi_j(\theta) \cong sE_iu_i''(c_i(\theta))\mathrm{Var}_i\phi_j(\theta)$, where we denote here $\psi_j(\theta) = \phi_j(\theta) - E_i\phi_j(\theta)$, $s = (1/n)g_j(z_j)$. For each state of nature θ we can approximate $u_i'(c_i(\theta))$ by a Taylor series around $s = 0$, i.e., around $\hat{c}_i(\theta) = c_i(\theta) - s\phi_j(\theta)$: $u_i'(c_i(\theta)) \cong u_i'(\hat{c}_i(\theta)) + u_i''(\hat{c}_i(\theta))s\phi_j(\theta) + u_i'''(\hat{c}_i(\theta))(s^2/2)\phi_j^2(\theta)$. Substituting from the similar Taylor series approximation of $u_i''(c_i(\theta))$, we then have for each θ $u_i'(c_i(\theta))\psi_j(\theta) \cong u_i'(\hat{c}_i(\theta))\psi_j(\theta) + u_i''(c_i(\theta))s\phi_j(\theta)\psi_j(\theta) - u_i'''(\hat{c}_i(\theta))(s^2/2)\phi_j^2(\theta)\psi_j(\theta)$. Under conditions which permit taking the expectation of the last expression (in particular if $\psi_j(\theta)$ is bounded), we then obtain our desired result. Here we have to use the independence of $\hat{c}_i(\theta)$ (and thus also of any function of it) from $\phi_j(\theta)$, and to use the smallness of s in order to neglect the last term which involves s^2.

be necessary. The conclusion reaffirms the Arrow-Lind argument in the dispute with Hirshleifer over the need for subsidization. Similarly, if there existed costlessly operating markets trading in the shares of all firms, then, due to the arguments given above, the excess social values η_j of each and every firm would be zero. In that case, too, private production would be socially optimal.

However, when the capital markets are deficient as assumed here in the farm model, the excess social values are not likely to vanish. In this case, it is clear by comparison of (13) and (4) that when taxation exists decentralized private investment would not be socially optimal. The reason that the market forces would fail to obtain the social optimum is that taxation introduces an element of externality. Through the distribution of the tax revenue, the tax collected from each private investor provides otherwise unavailable hedging to all other individuals. It is quite clear that investors are not likely to take into account this effect on others when making their own investment decision. Thus, some form of government intervention in private investment decisions will have to be employed in order to reach the "first best" social optimum feasible in the given market structure.

Among the various government actions possible to induce socially optimal private investment, we will consider here only the setting of differential rates of return. Indeed, it is quite obvious that if the central bank were to set a rate r_j for the transactions of firm j where:

$$(14) \qquad 1 + r_j = [1 + \rho][1 + t\eta_j]^{-1} \qquad j = 1, \ldots, n$$

then decentralized private investment would be optimal. Since for riskless firms the rate that would be set is ρ, we can think of that rate as the proper market rate of return.

To illustrate the use of the formula (14), as well as the general argument for subsidization, we return now to the example examined in the introduction of statistically independent alternative private and public projects. We now introduce a 50 percent tax rate.[13] In accordance with rule (8) the private project should then be evaluated using an average of the returns to the private investors and to the public at large. Since, due to the project's independence the latter return equals the expected return, the social return of the private project will be 1.075, an average of the 1.05 private risk-adjusted rate and the 1.1 expected rate. The private project will then be clearly superior to the public project whose social return is only 1.06. However, since the private entrepreneurs consider only the share of the project's benefits which accrues to themselves, they are not likely to undertake the project whose 5 percent net return to them falls short of the 6 percent return prevailing in the market. Thus, the government should subsidize the private project.[14] According to (12) we can compute $\eta_j = 1.1/1.05 - 1 \cong 0.0048$; and by (14), $1 + r_j = 1.06(1 + ½(0.048))^{-1} \cong 1.035$. Hence, if we assume that for the project size considered, the average return of the private project equals its marginal return, then formula (14) suggests that the government should encourage the private firm towards further investment by offering it loans at a subsidized rate of 3.5 percent.

In general, whether firm j will indeed face a subsidized rate of return for its borrowing (i.e., $r_j < \rho$) will depend on whether the excess social value of its output is positive. To analyze under what conditions the excess social value is positive we resort now to the use of a specific form of the subjective valuations, the form which would hold in mean variance analysis:[15]

[13]Note that if the tax rate were 100 percent, the optimal rule for the now-nationalized private firms would be the same as the rule (9) applicable for public firms.

[14]The data of the example can be interpreted as providing.

$$(1/\bar{y})f_k(\bar{y})E\phi_k(\theta) = 1.1;\ (1/\bar{y})f_k(\bar{y})S_k^k = 1.05$$

$$(1/\bar{y})g_k(\bar{v})\sum_{i=1}^{n}\tau_i S_k^i \cong (1/\bar{v})g_k(\bar{v})E\phi_k(\theta) = 1.06$$

$$r = \rho = 1.06$$

[15]It is shown by the author that the subjective valuation S_j^i will take the form of (15) if $u_i(c_i)$ is quadratic, if $c_i(\theta)$ *and* $\phi_j(\theta)$ are joint-normally distributed, or, as in (10) above, if $\phi_j(\theta)$ is only a small share in $c_i(\theta)$ and is independent of that remainder share of $c_i(\theta)$.

$$(15)\quad S_j^i = E_i\phi_j(\theta) - A_i cov_i(\phi_j(\theta), c_i(\theta)) \qquad i, j = 1, \ldots, n$$

where $A_i = A_i(c_i(\theta))$ was defined above in (11). If we further assume that all probability assessments are the same and that $\tau_i = 1/n$, $A_i = A$ for all i, we would get in the farm economy that:

$$(16)\quad S_j^j \eta_j = [1 - t]A\Big[f_j(y_j)\sigma_{jj} - \frac{1}{n}\sum_{i=1}^{n} f_i(y_i)\sigma_{ij}\Big]$$

where $\sigma_{ij} = cov(\phi_i(\theta), \phi_j(\theta))$. Thus if $\sigma_{ij} \leq 0$ for all $i \neq j$ we would clearly have $\eta_j > 0$; i.e., for a firm uncorrelated or negatively correlated to the rest of the economy, excess social value would be positive Moreover, from (16) it can be easily shown that the social value of the total private output exceeds the private evaluation of that output, i.e.:[16]

$$(17)\quad \sum_{j=1}^{n} f_i(y_j)\left[\sum_{i=1}^{n} \tau_i S_j^i - S_j^j\right] = \sum_{j=1}^{n} f_j(y_j) S_j^j \eta_j \geq 0$$

This result indicates that in some sense most η_j have to be positive. Negative excess social value should be regarded as the exception. For the excess social value of a firm j to be negative, according to (16) the firm would have to be highly correlated with most other firms, relatively inefficient (so that its "output" $f_j(y_j)$ *would be relatively low)* and also of a relatively low risk (i.e., a low σ_{ij}). The last consideration is due to the fact that the excess social value measures the divergence between social and private valuations; a divergence which is higher for private firms which are more risky and which therefore take larger risk margins than necessary from a social point of view.

The preceding analysis suggests that we should expect the social valuations to exceed the private valuations for most firms. That result would mean that for attainment of the social optimum most firms would have to face subsidized rates of return which are lower than the riskless rate.

IV. Second Best

The subsidization of the rates of return in the private sector was seen in the preceding section to facilitate the decentralization of the social optimum—the "first best" given the market structure. We now briefly consider the implication of possible government inability or unwillingness to charge different rates of return to different private firms. In that case the social planner seeking a "second best" would have to select levels of public investments z_j and a market rate of return r so as to maximize the social welfare function, given the overall resource constraint (5) and the fact that levels of private investments are set privately according to (4).

The intuitive results of this second best problem should be quite clear—the government ought to set the market rate of return below the shadow price of resources in order to encourage private investments. The technical solution of the second best becomes rather messy since we also have to take into account the effects of changes in the policy variables on private investments. To facilitate an easier exposition we thus assume once more than condition (7), which requires the equalization of the social value of marginal safe future consumption for all individuals, holds.

The first-order condition with regard to the optimal selection of the market interest rate can then be written as

$$(18)\quad \rho - r = [1 + r]t\left[\sum_{i=1}^{n} \frac{\partial y_i}{\partial r}\right]^{-1} \sum_{i=1}^{n} \eta_i \frac{\partial y_i}{\partial r}$$

[16]By use of (16),

$$\sum_{j=1}^{n} f_j(y_j) S_j \eta_j = [1 - t]A\left[\sum_{j=1}^{n} f_j^2(y_j)\sigma_{jj} - \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} f_i(y_j) f_i(y_i)\sigma_{ij}\right]$$

$$= [1 - t]A \sum_{j=1}^{n} \text{Var}[f_j(y_j)\phi_j(\theta) - \frac{1}{n}\sum_{i=1}^{n} f_i(y_i)\phi_i(\theta)] \geq 0$$

Given the structure of the farm model this term will be zero only if all firms were identical (and in particular in the same risk class) or if there were no aversion to risk in the economy, or if the tax rate was 100 percent

$$\equiv [1 + r]t\bar{\eta}$$

where $\bar{\eta}$ is an average of the various measures of excess social values η_i. If the presumption that most η_i are positive is correct, then $\bar{\eta}$ will probably be positive. Thus the market rate of return in the second best will have to be lower than the social rate of return.[17] Using (18), the rule for optimal public investment will be

$$(19) \quad g'_j(z_j) \sum_{i=1}^{n} \tau_i S^i_j = [1 + r][1 + t\bar{\eta}] + [1 + r]t\bar{\eta} \sum_{i=1}^{n} \left[1 - \frac{\eta_i}{\bar{\eta}}\right] \frac{\partial y_i}{\partial z_j} \qquad j = 1, \ldots, n$$

If we were to ignore the direct effects of public investment on private investments, the second best rule for public investment would indicate, on the one hand, use of the social valuation of output (generally higher than the private subjective valuation) as in the first best, and on the other hand, use of a safe rate of return which is equal to the social rate of return, but higher than the market rate of return. In order to evaluate the net result of the two effects on possible divergence of the public and private rules for investment, we can rewrite (19) (ignoring the effects $\partial y_i/\partial z_j$) as:

$$(20) \quad g'_j(z_j) S^j_j = [1 + r]\left[1 + \frac{t\bar{\eta} - \eta_j}{1 + \eta_j}\right] \qquad j = 1, \ldots, n$$

Thus all risk classes with excess social value higher than the average multiplied by the tax rate, the public rule for investment requires higher investment than would have been undertaken had the comparable private criteria for investment been adopted.

V. The Role of Taxation

The reader may wonder why the tax rate was not considered as one of the endogenous variables in the determination of the social optimum. Indeed, if we maximize social welfare with respect to the income tax rate t, the first-order condition which results requires (assuming that (7) holds) the equalization of the social and private valuations of the total output of the private sector. Under the simplifying assumptions which led to equation (17) above, this condition will necessitate a 100 percent tax rate, i.e., the nationalization of the entire private sector! Examination of this perplexing result will help to clarify the particular role of taxation in the farm model and the argument for subsidization.

The tax system in the farm economy does not provide any disincentives for efficient private investment and production. This characteristic of the model clarifies why a very high tax rate might be considered feasible; however, it does not explain why a high tax rate might be deemed desirable. To explain this result the exact nature of the income tax in the model must be examined. That tax system is basically a scheme of a negative income tax. It may alternatively be characterized as a form of mutual insurance[18] since every individual is required to pay a share of his risky profits into a common pool from which he then receives a share of the total sum collected. Since we assumed the absence of capital markets, this tax-insurance mechanism is the only means by which individuals in the farm model can obtain some hedge for their own risky enterprises. In the absence of any efficiency disincentives it is thus quite plausible that under some homogeneity conditions a full coverage, i.e., a 100 percent tax rate, will be socially desirable.

When the capital markets are perfect, it is quite clear that the insurance element in the tax scheme provides no improvement of the market's allocative mechanism. Indeed, in this case it can be shown that if initial lump sum redistribution of resources were possible, the social optimum could be reached without the use of the

[17] Also, if the government were able to borrow and lend in the international loan market at a uniform rate r^*, then ρ would be replaced by r^*

[18] Unlike the standard forms of insurance, in the present case of income taxation, "premiums" are paid only *after* the state of nature is revealed

now superfluous system of income taxation. In fact, however, capital markets even in the developed Western economies are not perfect. (In particular, insurance markets for future wage income are seriously deficient.) In this case our analysis indicated that the income tax structure can be thought of as providing some form of insurance. It is this aspect of insurance in the tax system which also creates the externality effect referred to above—one investor's taxed profits provide hedging for others.

Domar and Musgrave (and more recently Joseph Stiglitz and several others) have argued that income taxation is likely to lead by itself to increased private risk taking.[19] My analysis indicates that even where this result holds, when capital markets are imperfect, private risky investment should in general be further encouraged because of its positive external insurance effects. It should be emphasized that the above general argument—that income taxation provides a form of insurance and introduces an externality—does not require many of the assumptions which were made in the analysis for expositional purposes. In particular, the assumptions that lump sum initial redistribution is possible, that the whole tax revenue is returned to the public, or that it is returned according to preassigned shares, can be significantly relaxed.

VI. Conclusion

In the preceding analysis I have presented a case for the subsidization of risky private projects. This case was based on two key assumptions which hold in most countries: the existence of an operative income tax system and the incompleteness of the capital market structure. It is common knowledge that in reality a general practice of subsidization may create serious distortions and lead to abuse of public funds.[20] However, if considered as an alternative to direct public investment with its own known inefficiencies, subsidization of private projects may prove nevertheless to be the preferred practice.

[19]The above result on the equivalence of income taxation (with revenue redistribution) to initial lump sum redistribution when capital markets are perfect indicates that one should not expect any general increase in risk taking in this case. This conclusion points out a limitation of the partial equilibrium approach adopted by Domar and Musgrave (and Stiglitz) which ignores the use of the collected tax revenues.

[20]For a detailed empirical analysis of the practice of subsidization and its distortionary effects on firms which adopt "subsidy maximizing" policies, see Nachum Finger

REFERENCES

K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ Rev*., June 1970, *60*, 364–78.

M. J. Bailey and M. C. Jensen, "Risk and the Discount Rate for Public Investment," in Michael C. Jensen, ed., *Studies in the Theory of Capital Markets*, New York 1972.

P. A. Diamond, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, *57*, 759–76.

E. D. Domar and R. A. Musgrave, "Proportional Income Taxation and Risk Taking," *Quart. J. Econ.*, May 1944, *58*, 388–442.

Nachum Finger, *The Impact of Government Subsidies on Industrial Management: The Israeli Experience*, New York 1971

J. Hirshleifer, "Investment Decision Under Uncertainty: Applications of the State-Preference Approach," *Quart. J. Econ.*, May 1966, *60*, 252–77.

J. Mayshar, "Mean Variance Analysis: A Positive Approach," unpublished paper, Hebrew Univ. 1975.

A. Sandmo, "Discount Rates for Public Investment Under Uncertainty," *Int. Econ. Rev.*, June 1972, *13*, 287–302.

J. E. Stiglitz, "The Effects of Income, Wealth and Capital Gains Taxation on Risk Taking," *Quart. J. Econ.*, May 1969, *63*, 263–83.

"Strategic" Wage Goods, Prices, and Inequality

By JEFFREY G. WILLIAMSON*

In the long sweep of American experience, which income classes have benefited most by changes in relative consumer goods' prices? Have episodes of rapidly changing price structure coincided with dramatic shifts in earnings and income distributions? Have these trends and cycles reinforced each other so that *real* distribution indicators exhibit even more dramatic variation than conventional nominal indicators?

Obviously, inflation or deflation cannot have different expenditure effects by socioeconomic class unless *relative* prices of consumption goods exhibit some variance. This condition was apparently unfulfilled for the first two decades of postwar American experience. Robinson Hollister and John Palmer found that only medical care had changed significantly in relative price from 1947 to 1967. Prices of food, housing, clothing, transportation, personal care, and durables tended to conform closely to the overall consumer price index. In spite of a very wide range in budget shares from poor to rich, differential effects of postwar inflations have been relatively small on the expenditure side, at least prior to 1967. Hollister and Palmer concluded that relative consumer goods' prices had only a trivial influence on real distributions and that nominal distribution statistics were quite adequate social indicators. Lest this finding be applied indiscriminately to other historical episodes, it should be noted that American postwar growth has also been accompanied by remarkable stability in the distribution of nominal income (see Figure 1B). One wonders if the finding would hold prior to 1948, when inequality trends exhibited extraordinary variability (see Figure 1A), or even after 1967, when inequality was on the rise.

What was the distributional impact of relative price changes during the revolutionary income leveling following 1929, or during the wartime inflation following 1939? While this question has been central to modern British histories (see the contributions by Dudley Seers, Harold Lydall, and John Brittain), it has been ignored in America. Was the World War I inflation egalitarian? Who gained most from changing relative prices during the subsequent stabilization of the 1920's? Since these periods generated pronounced variation in the nominal earnings distribution, it should be of considerable interest to establish whether cost of living movements systematically reinforced or offset these inequality trends. Section I shows that since 1890 relative price movements have, with only one exception, always reinforced nominal distribution changes. Section III takes this analysis one step further and evaluates the quantitative importance of these cost of living effects on American size distributions. This is followed in Section IV by an attempt to identify, in John Muellbauer's (1974) language, those *strategic* commodities whose relative prices have influenced real expenditure distributions the most.

I. The Relative Prices of "Wage Goods"

Wage goods are defined here to include all consumption goods and services for which the income elasticity of demand is less than unity—that is, necessities. For the period 1914-48, urban price data are reported for three necessities (food, fuel and light, and rent) and three luxuries (clothing, house furnishings, and miscellaneous goods and services). The data prior to

*Professor, department of economics, University of Wisconsin, Madison The research reported here was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin-Madison by the Department of Health, Education, and Welfare pursuant to the Economic Opportunity Act of 1964 The opinions expressed are those of the author. The expert research assistance of Joan Hannon, David Ortmeyer, and James Roseberry is gratefully acknowledged An earlier and more detailed version of the paper (1975) is available upon request

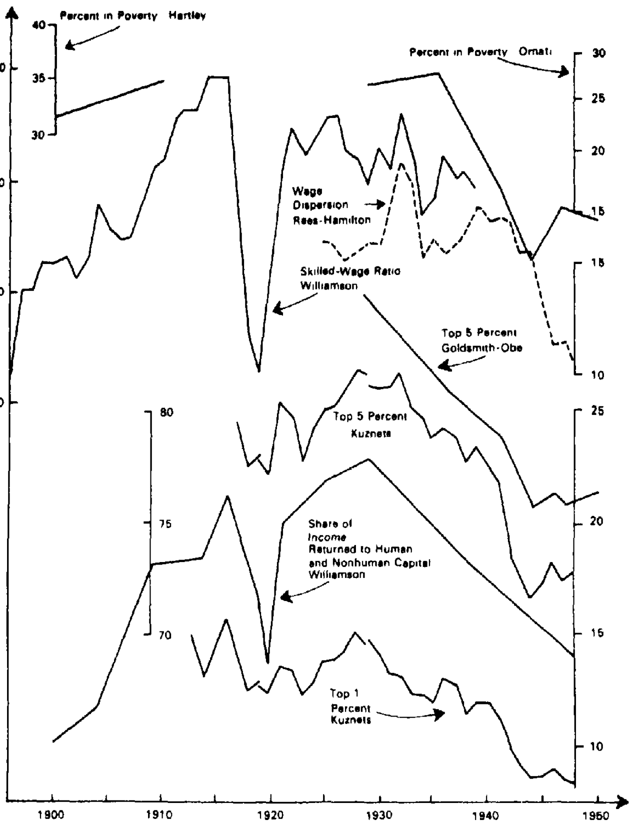


FIGURE 1A. TRENDS IN WAGE STRUCTURE AND INCOME DISTRIBUTION, 1896–1948
Source: the author, 1976a

1914 combine house furnishings with miscellaneous goods and services. The high level of aggregation is intentional, since it should help uncover any systematic price-distribution relationships should they be present, as well as facilitate comparisons with studies of other economies (see Muellbauer, 1973, 1974).

The evidence presented in Figure 1A documents a rise in nominal inequality from the mid-1890's to 1914. (See Peter Lindert and the author, 1976a, for more detailed evidence.) The peacetime inflation rate following 1896 was 2.1 percent yearly, modest by modern double-digit standards, but inflation nonetheless. Table 1 shows unambiguously that prices of necessities rose at a more rapid clip than did prices of luxuries. Indeed, while food prices rose by some 2.4 percent per annum in American cities, the prices of luxuries actually fell—and this without any quality adjustment during an age of revolution in consumer durables. Yet it is not the positive correlation between inflation on the one hand and a rise in the relative price of necessities on the other that deserves stress. Rather, it is the

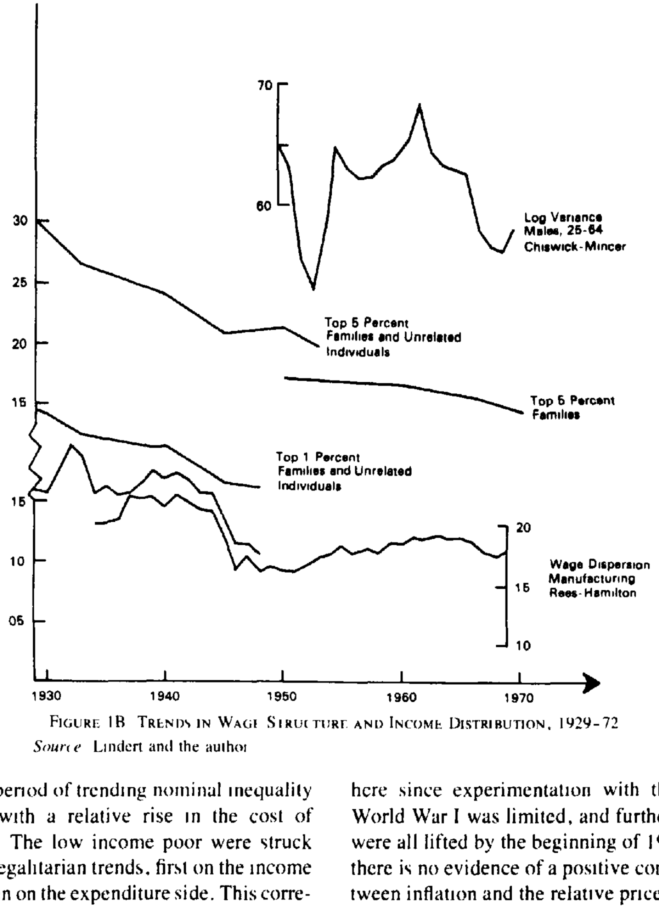


FIGURE 1B TRENDS IN WAGE STRUCTURE AND INCOME DISTRIBUTION, 1929-72
Source Lindert and the author

fact that a period of trending nominal inequality coincided with a relative rise in the cost of necessities. The low income poor were struck twice by inegalitarian trends, first on the income side and then on the expenditure side. This correlation can be found during the previous *century* of American economic experience, without exception (see the author, 1976b).

Turn now to the period following 1914, years for which both the nominal income distribution data and the urban price series are much improved. The rate of price inflation accelerates long before America enters the conflict, and the inflation continues to 1920: the annual rate was 14.6 percent over these six years. Yet, the urban price data in Table 2 show that luxuries rose in price far more dramatically than did necessities. (Price controls and rationing are not at issue here since experimentation with them during World War I was limited, and furthermore they were all lifted by the beginning of 1920.) While there is no evidence of a positive correlation between inflation and the relative price of necessities, there certainly is evidence of such a correlation between nominal inequality and the relative price of necessities. Figure 1A reveals a dramatic decline in every nominal inequality indicator from 1914 to 1920. It appears that relative price movements supported these nominal trends, in real terms, the egalitarian trend must have been considerably more dramatic.

Similar results emerge from an examination of the stabilization decade following 1920. The 1920's were years of trending inequality, so much so, judging from Figure 1A, that most if not all of the previous nominal egalitarian gains

TABLE 1—URBAN PRICE INDICES, 1890–1914

	Necessities			Luxuries	
	Food	Fuel and Light	Rent	Clothing	House Furnishings and Miscellaneous
1890	72	83	93	134	122
1891	72	86	93	135	119
1892	70	84	95	135	117
1893	72	84	95	128	114
1894	69	76	93	118	110
1895	68	78	90	113	103
1896	66	83	91	113	100
1897	68	80	88	110	96
1898	69	78	88	107	96
1899	70	79	87	106	95
1900	71	91	85	108	95
1901	74	92	87	103	93
1902	78	100	86	99	91
1903	77	112	91	98	93
1904	78	105	96	97	90
1905	78	101	97	96	87
1906	81	101	98	98	89
1907	85	101	102	102	96
1908	83	101	99	97	94
1909	84	100	97	95	95
1910	91	99	99	97	95
1911	93	95	97	96	96
1912	96	99	97	99	97
1913	97	102	100	101	98
1914	100	100	100	100	100

Note All prices (1914 = 100) are taken from Albert Rees, p 74

TABLE 2—URBAN PRICE INDICES, 1914–29

	Necessities			Luxuries			
	Food	Fuel and Light	Rent	Clothing	House Furnishings	Miscellaneous	Adjusted House Furnishings
1914	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1915	100.0	100.0	101.5	103.7	106.3	104.3	104.0
1916	120.0	107.3	102.3	118.8	122.9	110.0	117.8
1917	149.5	122.9	100.1	147.6	144.8	136.4	136.1
1918	178.4	144.6	105.3	211.3	197.1	158.5	182.1
1919	190.9	151.6	119.0	283.5	247.9	180.0	224.9
1920	174.6	190.1	142.5	268.4	267.6	199.8	237.5
1920	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1921	86.0	95.5	108.9	74.3	75.6	98.8	73.3
1922	83.6	98.4	109.5	65.8	72.5	95.9	68.5
1923	86.0	97.5	113.9	67.4	77.5	96.9	71.7
1924	86.0	96.0	116.2	65.5	74.6	97.1	67.4
1925	96.0	102.1	115.8	64.9	73.7	97.9	65.0
1926	93.4	99.7	114.2	63.7	71.4	98.2	61.5
1927	90.4	97.4	111.9	62.2	70.1	99.0	58.9
1928	89.2	96.5	109.1	61.7	68.2	99.6	56.0
1929	91.4	95.9	106.6	61.1	67.7	100.2	54.3

Note: See the author (1975, Table 2, p. 9) All prices are Bureau of Labor (*BLS*) estimates, except for Adjusted House Furnishings. The latter includes an estimated impact of the quality bias based on post-1935 evidence See Table 3

were lost. Table 2 shows once again that relative cost of living movements were supporting those nominal distribution trends. During this decade of stabilization, the price of luxuries declined far more precipitously than did that of necessities. Furthermore, the inequality bias in the changing price structure would be even more striking if we adjusted for quality improvements in consumer durables over the 1920's. In short, the poor must have found their relative economic position eroding from *both* the income *and* the expenditure side.

The striking association between distribution trends and changes in the commodity price structure does not continue through the distribution "revolution" between 1929 and 1948. This is not to say that the relative price structure of consumer goods and services was stable after 1929; Table 3 documents just the opposite. What is missing over the Great Depression decade is a consistent fall in the prices of *all* necessities relative to the prices of luxuries, although the characterization does hold for all commodities save one. The exception is important, however. The relative price of food rose between 1929 and 1948, as well as during the shorter term episode from 1936 to 1948. This represents a departure from a century of American experience (see the author, 1976b). Whether it is sufficiently striking to reverse the historical mutual reinforcement of expenditure and income effects on the size distribution is a matter reserved for Section III.

The postwar years can now be more adequately understood. The Hollister and Palmer finding of long-term stability in the structure of wage goods' prices (1947–67) is quite consistent with American twentieth century experience as a whole. When we discount war-induced cycles in unemployment and thus in the size distribution (see Charles Metcalf, and T. Paul Schultz, 1969), what remains in Figure 1B is only the weakest nominal egalitarian trend (see Schultz, 1971; Barry Chiswick and Jacob Mincer).

TABLE 3—URBAN PRICE INDICES, 1929–48

	Necessities			Luxuries				
	Food	Fuel and Light	Rent	Clothing	House Furnishings	Miscellaneous	Adjusted House Furnishings	Adjusted Miscellaneous
1929	130.8	112.3	146.7	118.1	116.0	106.0	136.8	111.3
1930	124.4	111.2	142.6	115.5	113.1	106.5	130.5	111.1
1931	102.6	108.7	135.2	105.1	101.8	105.5	114.5	109.3
1932	85.4	103.2	121.3	93.0	88.7	103.0	97.1	105.9
1933	83.0	99.8	104.5	90.1	87.4	99.7	93.4	101.8
1934	92.5	101.2	97.9	98.5	96.4	99.2	100.8	100.6
1935	99.1	100.5	97.7	99.2	98.4	99.4	100.7	100.1
1936	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1937	103.9	100.0	104.7	105.3	108.3	102.3	106.0	101.6
1938	96.5	99.7	108.0	104.7	107.3	102.8	102.6	101.4
1939	94.0	98.8	108.2	103.0	105.2	102.0	98.2	99.9
1940	95.4	99.5	108.5	104.2	104.4	102.4	95.2	99.6
1941	104.1	102.0	110.2	108.9	111.4	105.4	99.5	101.8
1942	122.3	105.2	112.6	127.4	126.9	112.4	108.3	107.8
1943	136.2	107.5	112.0	132.9	130.4	117.3	108.9	111.8
1944	134.4	109.6	112.2	142.2	141.6	122.9	115.7	116.4
1945	137.3	110.1	112.3	149.5	151.4	125.7	121.1	118.3
1946	157.6	112.2	112.7	164.1	165.3	130.5	129.4	122.0
1947	191.3	120.9	115.4	190.4	191.5	141.7	146.9	131.6
1948	207.5	133.6	121.8	202.9	203.3	151.9	152.7	140.2

Note: See the author (1975, Table 3, p. 11), and 1936 = 100. The Adjusted House Furnishings price index attempts to introduce a quality change estimate. The rate of quality improvement estimated for refrigerators (1935–48) is assumed to apply to all house furnishings over the period (Gordon 1971, Table 4, p. 144). The Adjusted Miscellaneous price index does the same for automobiles.

Stability in relative prices seems to correspond with stability in the size distribution of income. Or, if you like, stability in the output price structure coincides with stability in the input price structure—a very comforting result, but one that deserves far more attention and analysis than economists have given it.

What about the performance since the mid-late 1960's? Rates of inflation have accelerated, of course, but Table 4 shows once again that, with the trivial exception of personal care and miscellaneous items, necessities have risen in price more dramatically than luxuries. The size distribution of income has also deteriorated: Sheldon Danziger and Robert Plotnick have documented significant increased inequality in family and individual pretransfer nominal income from 1965 to 1972. History seems to be repeating itself—trending nominal inequality is exacerbated by changing prices of wage goods, which penalize the poor still further.

Now then, have these relative price movements served to influence the distribution of real incomes or expenditures in any significant way?

II. Prices and Inequality: Theory

Summary empirical measures of income distribution are essential to facilitate analysis and policy prescription, and many of them are exploited in Figures 1A and 1B. Shares of selected income classes, the variance, the coefficient of variation, the relative mean deviation, the standard deviation of logarithms, or the Gini coefficient are all arbitrary measures. All too frequently different statistics imply different conclusions regarding the behavior of inequality over time or in response to policy. This is so because each of these statistics implies some

TABLE 4—URBAN PRICE INDICES, 1967–75

	Necessities					
	Food	Fuel and Light	Housing	Medical Care	Personal Care	Miscellaneous (including tobacco and alcohol)
1967	100.0	100.0	100.0	100.0	100.0	100.0
1968	103.6	110.4	104.8	106.1	104.2	104.6
1969	110.6	112.9	113.3	113.4	109.3	109.1
1970	114.9	107.6	123.6	120.6	113.2	116.0
1971	118.4	115.0	128.8	128.4	116.8	120.9
1972	123.5	120.1	134.5	132.5	119.8	125.5
1973	141.4	126.9	140.7	137.7	125.2	129.0
1974	161.7	150.2	154.3	150.5	137.3	137.2
1975 (March)	171.3	163.0	166.6	164.6	148.9	146.5

	Luxuries						
	House Furnishings	Clothing	Auto	Other Transportation	Recreation and Education	Adjusted House Furnishings	Adjusted Automobiles
1967	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1968	104.4	105.4	103.0	104.6	104.7	102.6	105.4
1969	109.0	114.9	106.5	112.7	108.7	105.2	111.1
1970	113.4	116.1	112.7	128.5	113.4	107.4	119.3
1971	118.1	119.8	116.6	137.7	119.3	109.7	125.1
1972	121.0	122.3	117.5	143.4	122.8	110.4	128.2
1973	124.9	126.8	121.5	144.8	125.9	111.8	139.7
1974	140.5	136.2	136.6	148.0	133.8	122.5	154.5
1975 (March)	155.6	140.9	144.0	152.3	142.0	131.4	162.3

Note. See the author (1975, Table 4, p. 13). All prices (1967 = 100) are *BLS* estimates. The Adjusted House Furnishings price index attempts to introduce a quality change estimate, as does Adjusted Automobiles. The former is an estimate based on 1954–68 for refrigerators, and the latter is an estimate based on 1960–66 for "low-priced" autos (Gordon, 1971, Table 4, p. 144).

unique underlying social welfare function. A few years ago, Anthony Atkinson offered an ingenious device that confronts these issues with extraordinary economy. As Muellbauer (1973, 1974) has shown since, the Atkinson index is especially useful for confronting the distributional impact of inflation from the expenditure side.

Among the statistics listed above, those used most commonly are the coefficient of variation, the relative mean deviation, the Gini coefficient, and the standard deviation (or variance) of logarithms. The most recent and influential American distribution studies, for example, have been those by Schultz, Chiswick and Mincer, all of whom used *log* variance statistics. Although the *log* variance statistic may be an attractive extension of the human capital model, it implies a special degree of "political aversion" to economic inequality.[1] Since the statistic is constructed relative to the mean, it follows that equiproportional growth implies constant inequality over time. Atkinson (p. 257) proposed a more general index that has this property:

$$(1) \quad I = 1 - \left\{ \sum_j \left[\frac{\bar{y}_j}{\bar{y}} \right]^{1-\epsilon} f(y_j) \right\}^{1/(1-\epsilon)}$$

where $0 \leqq I \leqq 1$: $\bar{y}_j$ is mean income of the jth class; $\bar{y}$ is mean income, economy wide; $f(y_j)$ is percent in the jth class; and $\epsilon \geqq 0$ is a parameter measuring the *degree* of inequality aversion. This is a very attractive general statistic, since it allows us to examine inequality experience while applying various weights to the relative importance of different intraclass transfers. As ϵ rises, transfers to lower income groups are given heavier weight and transfers among top income recipients are given lighter weight. Consider two extremes. If $\epsilon = 0$, we are in effect describing a society in which only aggregate growth counts and distribution is irrelevant (or, less harshly, policymakers rely completely on "trickling down"). In this case, $I(t) = 0$ for all t. As $\epsilon \to \infty$, society tends to take greater account of transfers to the poverty group and ignores the source of the transfers—the distribution of income among the "nonpoor" has no political relevance. In the empirical analysis which follows, we shall consider values of ϵ in the range 1.5 to 4.0. Based on Atkinson's experiments (pp. 260–62) with United States 1950 size distribution data, this range encompasses such popular distribution statistics as the Gini coefficient and the standard deviation of logarithms.

The algebraic elegance of Atkinson's index should not deceive us into believing that ϵ is independent of the measured inequality. On the contrary, ϵ itself is an endogenous variable that should certainly have risen following 1929. Increasing equality, however measured, ensures greater political participation by lower income groups. Increasing relative political participation of the poor implies less legislative tolerance for inequality and a rising ϵ. Having made this confession of endogeneity, however, we fail in this paper to supply any useful resolution. We shall instead treat ϵ as an exogenously given parameter and explore the implications of its size.

It should be apparent that Atkinson's index can be a very powerful tool in evaluating the impact of competing policies, which may have complex and uneven effects by income class. It also supplies a means by which changes in the distribution can be interpreted. The index can be rewritten (see Atkinson, p. 250) as

$$(2) \qquad I = \frac{\bar{y} - y_E}{\bar{y}}$$

To be consistent with the empirical work that follows in Section III, y will denote consumption expenditures rather than income. Thus, y_E is the per capita consumption level which if given to everyone would generate the same aggregate welfare as the current consumption expenditure distribution. This "equally distributed equivalent level of consumption" is derived from a social welfare function in which ϵ is a parameter. The social welfare

[1]The word "special" should not imply political irrelevance, judged by the fact that Chiswick's *log* variance statistic appears in the 1974 *Economic Report of the President*

function has the properties of constant relative inequality-aversion, a property implied by Atkinson's index in expression (1). Space precludes further discussion of this point (see Atkinson), but perhaps some examples may prove helpful. If under some assumed value of $\epsilon, I = 0.20$, then the index tells us that the same level of aggregate welfare could be achieved by distributing equally only 80 percent of aggregate consumption—"essential consumption" in Paul Baran's words. Once again in Baran's words, 20 percent of aggregate consumption would be "potential economic surplus."[2] Alternatively, if a given policy lowers I from .21 to .20, the increase in social welfare would be an increase of 1 percent in equally distributed consumption. The reader will also note the following: Two very different assumed values of ϵ, say 1.5 and 4.0, may imply two very different initial levels of inequality, say $I_{1.5} = 0.20$ and $I_{4.0} = 0.70$, while the proposed policy may raise them both by approximately one percentage point—a 1 percent decline in equality distributed consumption in both cases. It seems to me that such calculations have far more intuitive appeal in judging what is "large" and what is "small" than do changes in the Gini coefficient, the *log* variance statistic, or even numbers in poverty.

Now then, can we improve on our measures of the expenditure side incidence of inflation? Let us define a new index, $\hat{I}$, which deals with prices from the expenditure side. Our real inequality index will be written as

$$(3)\quad \hat{I} = 1 - \left\{ \sum_j \left[\frac{\bar{y}_j^*}{\bar{y}^*} \right]^{1-\epsilon} f(y_j) \right\}^{1/(1-\epsilon)}$$

where $\bar{y}_j^* = \dfrac{\bar{y}_j}{\sum_j w_{ij} p_{ij}}$ = mean real expenditure of the jth income class

$$\bar{y}^* = \sum_j \left\{ \frac{\bar{y}_j}{\sum_j w_{ij} p_{ij}} \right\} f(y_j)$$

= economy wide mean real expenditure

and w_{ij} = fixed expenditure share on the ith good in the jth income class. Muellbauer (1974, pp. 38-42) uses the linear expenditure system to estimate variable budget weights by expenditure class, an attractive procedure since it permits consumers to substitute one commodity for another in response to relative price changes. Our approach will be more primitive, since we shall utilize fixed (Paasche) budget weights estimated from Engel functions. While our implicit price indices by income class fail to satisfy the constant utility criterion of Fisher's true cost of living index, they do have the great advantage of computational simplicity. Furthermore, recent research by Muellbauer (1973), Marilyn Manser and Laurits Christensen has suggested that in practice little is lost by our inelegant econometric strategy.

Given the distribution of nominal expenditures, how do changing relative prices on the expenditure side affect the distribution of real consumption levels? There are four component parts to the answer, and the congruence of these four forces would ensure a potent impact of prices on twentieth century inequality experience. First, are the Engel functions sufficiently steep so that wide variances in budget weights are observed over income classes? Second, is there a wide variance in total family consumption by income class? Obviously, if nominal income is equally distributed, expenditure patterns are likely to exhibit very little variation, and thus relative price changes will affect all families equally. On these grounds alone, a given rise in the relative price of, say, foodstuffs is bound to have a greater inegalitarian impact in a society with great nominal inequality to begin with. Third, relative prices themselves must change. Fourth, relative prices within the necessity and luxury categories must behave consistently Section I documented the required consistency for almost every period since the turn of the century, the exception being food

[2] The remainder of *GNP* would be going to gross investment to insure that current welfare levels are sustainable Baran would also allocate to potential surplus that portion of *GNP* going to defense regrettables (see William Nordhaus and James Tobin).

prices during the Great Depression and World War II.

What, then, has been the historical impact of prices on inequality in twentieth century America?

III. Prices and Inequality: Fact

A. *Lessons from an Earlier Era of Instability, 1917–29*

Compared with the postwar episode following 1948, the first third of the twentieth century was a period of extraordinary volatility in income distribution, relative prices, and output mix. How important was the relative price structure as an influence on distribution from the expenditure side? The answer may be relevant for the 1970's, another period of structural volatility. Unfortunately, adequate size distribution data (urban or economy wide) do not become available until 1935–36. However, there is no reason why we can't set aside Atkinson's index for the moment and construct instead some primitive measures using Kuznets's top 10 and 5 percent bands. The computations are reported in Table 5.

TABLE 5—DEFLATION OF RELATIVE INCOME SHARES, NONFARM, 1917–29
(in percent)

Year	Top 10 percent Money	Top 10 percent Real	Top 5 percent Money	Top 5 percent Real
1917	34.5	34.5	25.6	25.6
1920	30.3	29.3	22.6	21.8
1920	30.3	30.3	22.6	22.6
1929	35.4	36.3	26.1	26.8

Note: See the author (1975, Table 5, p. 23). The nominal income shares are Kuznets's nonfarm estimates. The real income shares are derived by using computed cost of living indices. The commodity price data are taken from Table 2 and the weights are estimated from double logarithmic expenditure functions, 1918–1919 urban survey data.

It appears that as much as a fifth of the 5.2 percentage point decline in the top 10 percent share can be explained by relative cost of living movements from 1917 to 1920. Similar results are apparent for the 1920's: While the top 10 percent share in real income rose by 6 percentage points, one of these percentage points was due to the favorable cost of living changes facing the rich.

TABLE 6—PRICES AND URBAN INEQUALITY, 1914–20
(Using 1935—36 Weights)

Atkinson's Inequality Index:	$\epsilon = +1.5$	$\epsilon = +2.5$	$\epsilon = +4.0$
Nominal expenditure distribution (urban, 1935–36)	1706	.2592	.3574
Impact of historic price changes, 1914–20 (using 1935–36 weights)			
All prices	.1602	2447	3399
Food	.1868	2814	.3836
Rent	1736	.2631	3618
Fuel and light	.1785	2703	.3710
Furnishings	1665	.2534	.3504
Clothing	1576	.2409	3351
Miscellaneous	1465	.2253	.3160

Note: The nominal distribution index is for urban families, 1935–36, as are the expenditure weights. See Table 7 for sources and methods. The prices used in the calculations are from Table 2.

It's a pity that urban size distribution data are unavailable for this period, since the reader may be skeptical about the relevance of a calculation based only on the top 5 or 10 percent. An alternative device for establishing the distributional impact of relative price changes during the pre-1929 period is presented in Table 6. Here we ask: What would have been the impact of the 1914–20 relative price changes on the 1935–36 distribution of (real) expenditures among urban families? Three values of ϵ are used, and they tell a consistent story. That is, those relative price changes served to lower the incidence of urban inequality from 1 to 1.7 percentage points. The measured influence would be far greater, of course, were we able to take account of farm families as well.

B. *The Egalitarian "Revolution," 1929–48*

Table 7 documents the impact of prices on inequality during the 1930's and 1940's, a period of impressive nominal leveling in the distribution. Look first at the longer term, the two decades from one full-employment year to another, 1929–48. Relative prices did tend to contribute to the egalitarian drift. The impact is

TABLE 7—PRICES AND URBAN INEQUALITY: 1929–48

Atkinson's Inequality Index:	$\epsilon = +1.5$	$\epsilon = +2.5$	$\epsilon = +4.0$
Nominal expenditure distribution (urban, 1935–36)	1706	2592	3574
Impact of historic price changes, 1935–48			
All prices	.1735	2632	3622
Food	.1922	2887	3921
Housing	1722	2613	3598
Fuel and light	.1737	2635	3627
Furnishings	1681	.2556	3530
Clothing	1622	2473	3430
Miscellaneous	1565	.2395	3336
All prices: quality adjusted	1766	.2676	.3674
Furnishings	.1693	2573	3551
Miscellaneous	.1593	.2435	.3385
Nominal expenditure distribution (urban, 1935–36)	1706	2592	3574
Impact of historic price changes, 1929–48			
All prices	.1662	2532	3504
Food	1838	2774	3790
Housing	.1693	2575	3554
Fuel and light	.1724	2617	3604
Furnishings	.1687	2565	3542
Clothing	.1646	2507	.3471
Miscellaneous	1586	2424	3372
All prices: quality adjusted	1714	2604	3590
Furnishings	1703	2588	3569
Miscellaneous	1630	.2487	.3448

Note: See the author (1975, Table 7, p 26) The nominal distribution indices are for urban families, 1935–36 These data refer to total family expenditures over twelve income classes ranging from \$0–\$500 (excluding those on relief) to \$5000–\$10,000 The calculation in line 1 uses expression (1) in the text The impact of prices on inequality uses line 1 as a base, following expression (3) in the text The p_{ij} are taken from Table 3 and the w_{ij} are estimated from double-log expenditure functions.

hardly as great as during the more volatile first third of the twentieth century: Atkinson's index is lowered by only 0.4 to 0.7 percentage points in response to the price changes, and the index falls hardly at all when estimates of quality improvements are introduced. Nevertheless, nominal and real distribution once again move alike.

Curiously enough, the correspondence fails for the shorter term period following 1935. The rise in food prices (Table 3) was sufficiently large to reverse the correlation: While nominal inequality indicators were falling (primarily in response to full-employment effects, according to Schultz, 1971, Chiswick and Mincer), living costs were rising most dramatically for the poor—the *only* such correspondence in a century of American experience. It should be emphasized, however, that the more relevant long-term experience, between the full-employment points 1929 and 1948, yields the more conventional result: While nominal inequality indicators are falling, living costs were also rising most dramatically for the rich.

C. *Prices and Inequality in the 1970's: A Recurring Theme*

Income inequality was on the rise between 1967 and 1973. No doubt expenditure distributions would exhibit a less steep inequality trend, but the income distribution data in Table 8 are adequate to gauge the impact of prices on urban inequality. Table 9 confirms that relative price movements were contributing significantly to the inequality trend. (Although their primary focus is on the "poor," similar results can be found in publications by Thad Mirer, and Robert Plotnick and Felicity Skidmore, p. 126.) Assuming $\epsilon = 2.5$, the historical price changes from 1967 to March 1975 would have raised urban inequality from a .3165 base (the 1960–61 figure) to .3298, a rise of 1 33 percentage points. This is no small matter when judged by the actual increase in nominal income inequality from 1967 to 1973. At $\epsilon = 2.5$, Atkinson's

TABLE 8—TRENDS IN NOMINAL INEQUALITY, 1967–73

Atkinson's Inequality Index	$\epsilon = +1.5$	$\epsilon = +2.5$	$\epsilon = +4.0$
1967	3274	5342	7092
1968	3157	5309	7190
1969	.3222	5445	.7349
1970	.3411	.5824	.7714
1971	.3324	5736	.7701
1972	3369	.5824	.7796
1973	3328	.5842	.7902

Note: See the author (1975, Table 9, p. 29). The data refer to family incomes. To make the data conform to the requirements of Atkinson's index, the bottom three income classes have been collapsed to one, \$1999 and below.

TABLE 9—PRICES AND URBAN INEQUALITY: 1967–75

Atkinson's Inequality Index:	$\epsilon = +1.5$	$\epsilon = +2.5$	$\epsilon = +4.0$
Nominal expenditure distribution (urban, 1960–61)	.1913	.3165	.4618
Impact of historic price changes, 1967–March 1975			
All prices	.2004	.3298	.4773
Food	2178	.3573	.5097
Housing	.1944	.3206	.4659
Fuel and light	.1924	.3180	4633
Furnishings	.1889	.3125	4565
Clothing	.1869	.3097	4536
Miscellaneous	1780	.2959	.4367
All prices: quality adjusted	1982	.3273	4748
Furnishings	1899	.3142	.4587
Miscellaneous	1752	.2918	.4320

Note See the author (1975, Table 8, p 28). The nominal distribution indices are for urban families, 1960–61 These data refer to total family expenditures over ten income classes ranging from "under $1000" to "$15,000 and over" The calculation in line 1 uses expression (1) in the text The impact of prices on inequality uses line 1 as a base following expression (3) in the text The p_{ij} are taken from Table 4 and the w_{ij} are estimated from double-log expenditure functions

index rises by 5 percentage points. That is, price trends have had at least one quarter as much effect as nominal income trends in contributing to inequality movements in recent years.

We conclude that prices have been a significant regressive force since the late 1960's.

IV. "Strategic" Wage Goods and Inequality

Which wage goods have been most responsible for the regressive impact of prices since 1967? Table 9 confirms in quantitative terms what we already suspected. Virtually all of the regressive price impact can be traced to food. While overall inflation acted to raise Atkinson's index ($\epsilon = 2.5$) by 1.33 percentage points, food prices by themselves contributed to a 4.08 percentage point increase. Similar patterns emerge for the 1929–48 period (Table 7). The total impact of the most recent inflation was less regressive primarily because of the rise in price of all goods contained in the "Miscellaneous" category: automobiles, medical care, education, and recreation. Nonetheless, food prices do indeed dominate, and price regressivity prevails after 1967.

Certainly food's strategic role is explained in part by its extraordinary rise after 1939 in the first inflationary episode, and after 1972 in the second. But its dominance can also be explained by the relative sensitivity of inequality measures to a given change in food prices in comparison with an identical change in any other price. The computations reported in Table 10 reveal the impact of a 25 percent change in some consumer good price—holding all other prices constant—on Atkinson's index. The strategic wage goods are food, whose price has by far the largest potential regressive impact, and miscellaneous, whose price has by far the largest potential progressive impact. The latter includes car purchases and general services, both of which have an important progressive price impact.

Some surprises emerge when these computed "price-sensitivity" figures are compared with figures from other countries or from earlier points in American history. Food is a far more important strategic wage good in America than it is in the United Kingdom. Muellbauer's estimates (1974, Table 11, col. 1) for the United Kingdom are not completely comparable with ours, yet the American figure is *twice* as large. Table 10 also reports the surprising result that food prices have *increased* their strategic role since the 1930's. In 1935–36, a 25 percent rise in food prices would have raised Atkinson's index ($\epsilon = 1.5$) by 0.63 percentage points; the comparable figure for 1960–61 is 1.03 percentage points. There are some other changes over these three decades that are worth noting too: (i) fuel and light have declined from a position of important price regressivity to unimportance; (ii) medical care has reversed its role from price progressivity to price regressivity.

V. Qualifications and Speculations

The sensitivity of distribution indices to prices in twentieth century America is in part a fabrication—but, I believe only a small part. A fixed-budget weight model has been used throughout. More elaborate expenditure models, with endog-

TABLE 10—SENSITIVITY ANALYSIS, "STRATEGIC" COMMODITIES, INFLATION, AND INEQUALITY

Atkinson's Inequality Index.	*U.K.* All Families 1970 $\epsilon = +1.5$	*U.S.* Urban Families 1935–36 $\epsilon = +1.5$	1960–1961 $\epsilon = +1.5$	1960–1961 $\epsilon = +2.5$	1960–1961 $\epsilon = +4.0$
Nominal expenditure distribution I	0962	1706	1913	3165	.4618
Impact of a 25 percent change in P_j, holding all other P_k constant, $I - \hat{I}$					
Detail, $j = 1 \ldots 6$					
Food	+ 0054	+.0062	+.0099	+ 0156	+ 0190
Housing	+ 0028	+ 0018	+ 0011	+.0015	+ 0015
Fuel and light	0	+ 0023	+ 0004	+ 0006	+.0006
Furnishings	na	− 0006	− 0007	− 0011	− 0014
Clothing	− 0013	− 0022	− 0028	− 0042	−.0051
Miscellaneous	na	− 0073	−.0077	−.0122	− 0150
Detail, $j = 1, \ldots, 15$					
Food	+ 0054	+ 0063	+ 0103	+ 0161	+.0194
Housing	+ 0028	+ 0019	+ 0012	+ 0016	+ 0016
Fuel and light	0	+ 0023	+ 0004	+ 0006	+.0006
Furnishings	} − 0043	− 0006	−.0007	−.0011	− 0014
Automobile		− 0041	− 0053	− 0078	− 0088
Household operations	na	− 0018	− 0004	− 0008	− 0011
Clothing	−.0013	− 0021	− 0027	− 0043	− 0051
Other transport	na	0	− 0001	− 0003	− 0004
Medical care	na	− 0004	+ 0007	+ 0009	+ 0010
Recreation	− 0008	− 0012	− 0011	− 0017	− 0020
Personal care	na	+ 0001	0	0	− 0001
Tobacco	+ 0012	+.0003	+ 0001	+ 0002	+ 0002
Education	na	− 0005	− 0004	− 0006	− 0007
Reading	na	+ 0001	0	0	0
Miscellaneous goods	− 0011	} 0	} − 0025	} − 0039	} −.0046
Miscellaneous services	+ 0016				

Note: See the author (1975, Tables 10 and 11, pp.32–33) and Tables 7 and 9 above. The 1970 *U.K.* data come from Muellbauer (1974, Table V, p. 47) and refer to the distribution of 1970 expenditures in 1964 prices.

enous budget weights, would diminish our price-sensitivity estimates. Since food is the overwhelming strategic wage good, however, it appears unlikely that approaches allowing for substitution would significantly change our results.

While this paper establishes that nominal and real inequality indicators almost always move together, the differential impact of prices by class has never been as large as the nominal inequality movements themselves. One can hardly dismiss them on these grounds, however, since any serious macro-economic distribution theory must confront these curious price facts. Why the consistent historical correlation between the relative prices of outputs and inputs? That is, why do periods of "stretching" in the pay structure and increasing nominal inequality in the size distribution *always* contain relative price changes which inflate the cost of living for the poor faster than for the rich? Why is the opposite *almost always* true for periods of levelling in the nominal size distribution? The correspondence seems to extend to periods far longer than can be accounted for by aggregate demand instability. It seems to me that any macrodistribution theory which purports to explain American twentieth century inequality experience must simultaneously account for the behavior in the commodity output price structure as well.

REFERENCES

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

Paul A. Baran, *The Political Economy of Growth*, New York 1957.

J. A. Brittain, "Some Neglected Features of Britain's Income Levelling," *Amer. Econ. Rev. Proc.*, May 1960, *50*, 593–603.

Barry R. Chiswick, *Income Inequality*, New York 1974.

——— **and J. Mincer,** "Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985," *J. Polit. Econ.*, May/June 1972, *80*, Part II, S34–S66.

S. Danziger and R. D. Plotnick, Demographic Change, Government Transfers, and the Distribution of Income," disc. pap. no. 274–75, Inst. Res. Poverty, Univ. Wisconsin 1975.

R. J. Gordon, "Measurement Bias in Price Indexes for Capital Goods," *Rev. Income and Wealth*, June 1971, *17*, 121–73.

R. Hollister and J. Palmer, "The Impact of Inflation on the Poor," disc. pap. no. 40–69, Inst Res. Poverty, Univ. Wisconsin 1969

Simon Kuznets, *Shares of Upper Income Groups in Income and Savings*, New York 1953.

P. H. Lindert and J. G. Williamson, "Three Centuries of American Inequality," disc. pap. no. 333–76, Inst. Res. Poverty, Univ. Wisconsin 1976.

H. F. Lydall, "The Long-Term Trend in the Size Distribution of Income," *J. Royal Statist. Soc.*, Part I, 1959, *21*, 1–37.

M. Manser and L. Christensen, "Cost of Living Indexes and Price Indexes for U.S. Meat and Produce, 1947–1971," S.S.R.I., Univ. Wisconsin, Oct. 1973.

Charles E. Metcalf, *An Econometric Model of the Income Distribution*, Chicago 1972.

Jacob Mincer, *Schooling, Experience and Earnings*, New York 1974.

J. Muellbauer, "Prices and Inequality: The Recent U.K. Experience," disc. pap. no. 6, Univ. London, Birkbeck College, July 1973.

———, "Prices and Inequality: The United Kingdom Experience," *Econ. J.*, Mar. 1974, *84*, 32–55.

W. Nordhaus and J. Tobin, "Is Growth Obsolete?," in *Economic Growth: Fiftieth Anniversary Colloquium V*, New York 1972.

Robert D. Plotnick and Felicity Skidmore, *Progress Against Poverty*, New York 1975.

Albert Rees, *Real Wages in Manufacturing, 1890–1914*, New York 1961.

T. P. Schultz, "Secular Trends and Cyclical Behavior of Income Distribution in the United States, 1944–1965," in Lee Soltow, ed., *Six Papers on the Size Distribution of Wealth and Income*, New York 1969.

———, "Long Term Change in Personal Income Distribution: Theoretical Approaches, Evidence and Explanations," Rand Corporation, Santa Monica, Nov. 1971.

Dudley Seers, *Changes in the Cost-of-Living and the Distribution of Income Since 1938*, Oxford 1949.

J. G. Williamson, "Strategic Wage Goods, Prices and Inequality," disc. pap. no. 294–75, Inst. Res. Poverty, Univ. Wisconsin 1975

———, (1976a) "The Sources of American Inequality, 1896–1948," *Rev. Econ. Statist.*, Nov. 1976, *58*, 387–97.

———, (1976b) "American Prices and Urban Inequality Since 1829," *J. Econ. Hist.*, June 1976, *36*, 303–33.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1974.

Inequality: Earnings vs. Human Wealth

By LEE A. LILLARD*

Inequality in income has been an issue of continuing interest to economists. The recent literature on earnings over the life cycle (see Yoram Ben-Porath, Thomas Johnson, William Haley) has made clear that a relevant policy issue is the inequality of lifetime earnings appropriately discounted, or human wealth To the extent that earnings differences among individuals at one point in time are compensated at another, the equity issue is diminished. For example, among recent entrants into the labor force low earnings may reflect high levels of investment in training which will be compensated by high earnings at later stages of the life cycle, and conversely for those investing less. On the other hand, earnings differences may reflect permanent differences in earnings capacity which persist throughout the life cycle. In cross-section data it is not possible to distinguish compensatory from permanent differences among individuals except when due to measurable variables. Indeed one cannot distinguish these two from purely transitory variation. In this paper[1] I estimate inequality in human wealth, evaluated from ages 16 to 65, from data for a group of men born between 1917 and 1925 (a birth cohort) for whom earnings data are available at several points in their life cycle. I also examine schooling, ability, and background as determinants of human wealth inequality.

This analysis of longitudinal data extends previous studies of earnings and income inequality. Most previous studies considered earnings differences among members of a population at a point in time (a cross section), although comparisons have been made among demographic groups over time. Attempts by researchers to incorporate the life cycle notion include 1) calculation of inequality within narrow age groups, and 2) calculation of the present value of the cross-sectional lifetime profile of earnings. For example, see Hendrik Houthakker and Bruce Wilkinson. Similarly, Morton Paglin recently suggested that measured inequality should include only variation around the aggregate cross-sectional[2] mean earnings-age profile.

Longitudinal data allow separation of earnings into permanent and transitory components. While permanent differences among individuals due to measurable variables such as years of schooling and socioeconomic background can be ascertained from either cross-section or longitudinal data, multiple observation of each individual is necessary to estimate the variation in permanent earnings differences among observationally identical individuals. A logical use of this additional information is to consider inequality among individuals in the discounted sum of lifetime earnings.

I conclude from the following analysis of longitudinal data (and from other studies by the author cited at appropriate places) that there is considerable variation in human wealth. Human wealth is however substantially more equally distributed, as measured by the coefficient of variation or Gini coefficient, than earnings within narrow life cycle ranges. The

*Research associate, National Bureau of Economic Research. The work for this paper was supported by the National Science Foundation (GS31334) and the U S Department of Labor (L73-135). I wish to thank Finis Welch and Victor Fuchs for their helpful suggestions, and Barbara Williams for her assistance

[1]A similar approach is taken in an earlier paper by the author (1975) where the emphasis is on the usefulness of empirical earnings functions for generating earnings and human wealth distributions The paper provides a more detailed account of human wealth as an index of lifetime economic well-being.

[2]The Paglin procedure corrects only for the overall mean age-income relationship and ignores individual differences in profiles which are predictable, such as schooling differences His procedure would indicate positive inequality even if all individual or household age-income profiles were equal in present value but differed in shape from the aggregate mean profile.

contribution of schooling, measured cognitive ability, and a limited set of background variables to variation in human wealth is approximately the same as their contribution to variation in earnings within age groups. Both are roughly 10 to 12 percent. The remaining inequality in human wealth is due to individual differences in earnings which persist over a lifetime caused by unmeasured factors. Of the measured variables, years of schooling has a larger effect on annual earnings than does measured ability, at any age. However, the contribution of schooling to human wealth is much more sensitive to discounting because of the period of foregone earnings associated with schooling. Consequently, measured ability has a substantial effect on human wealth that persists even at discount rates sufficiently high to make the return to schooling negative.

I. Earnings, Human Wealth, and the Life Cycle

It may be useful in developing a clear understanding of the relationship between variations in human wealth and annual earnings to consider some straightforward illustrations. For simplicity, assume away exogenous earnings growth over time and differences in the length of working life and retirement. Also assume a zero discount rate so that present values are sums. Consider first an example which should clarify how the variance in earnings can exceed the variance in human wealth. Assume that all individuals in a given population have the same lifetime earnings profile (earnings rise with age), but they differ in age There is correspondingly zero variation in human wealth and zero variation in earnings at any specified age, but positive variation in earnings at a point in time among individuals in the population. In addition, one would observe positive covariance among earnings values for adjacent years. Those with high earnings in the first year are older and have high earnings in the second year as well. If there are several earnings streams which have the same present value, but differ in the rate of earnings growth with age, these conclusions are unaltered except that there will be positive variance in the earnings of individuals the same age, and that earnings early in life and late in life will be negatively correlated among individuals. Clearly in this illustration the coefficient of variation and Gini coefficient of concentration will indicate zero inequality in human wealth and positive inequality in earnings, by age group, or aggregated over ages.

Secondly consider how the variance in human wealth can exceed the variance in earnings. Assume contrary to the first illustration that earnings do not vary with age (flat age-earnings profiles), but do vary among individuals. Variation in earnings will be the same at all ages and aggregated over ages. The variance in human wealth *must* exceed the variance in earnings since it is the discounted sum of the constant earnings value. In this particular case the coefficient of variation in earnings at any age exactly equals the coefficient of variation in human wealth. Inequality in earnings would then appropriately index inequality in human wealth. If the age-earnings profiles are allowed to slope upward (but remain parallel), any cross-sectional earnings distribution aggregating over ages will have larger variance than the earnings distribution at any age. The variance of the aggregate earnings distribution will depend on the age distribution of members of the aggregate as well as the distribution of profiles among members. The inequality in earnings at any age still accurately reflects inequality in human wealth even though the variance in human wealth is larger than the variance in earnings at any age.

Clearly as the features of these two extreme illustrations are combined, i.e., as individual profiles differ in both mean level and lifetime pattern, either extreme may dominate. The major difference between the two illustrations is the degree to which differences in lifetime earnings profiles among individuals are compensated or uncompensated in present value and the degree of variation in uncompensated differences. The model to be considered empirically incorporates all of these features: various shapes of age-earnings profiles due to measur-

able variables, differences in human wealth dur to unmeasured variables, as well as stochastic variation in earnings from year to year.

II. *Ex Post* Pattern of Lifetime Earnings

In this section the lifetime pattern of real annual earnings[3] (in 1970 dollars) is analyzed for a group of 4699 men for whom 2 to 5 age-earnings points are observed between ages 19 and 57, and years between 1943 and 1970. These are the men in the *National Bureau of Economic Research-Th* sample.[4] The sample is based on a group of males born primarily between 1917 and 1925 and volunteering for Air Force pilot, navigator, and bombardier programs in 1943. The men are all in the top half of *AFQT* exam scores, are at least high school graduates and have experienced military service (and thus were eligible for G.I. Bill benefits). In addition to annual earnings and schooling, each individual reported several background characteristics including parents' education, number of siblings, religion, and number of childhood family moves. Several measures of ability taken upon entry into the military have been aggregated into a single index of ability corresponding roughly to *IQ*.[5]

The basic data on individual schooling, *IQ* type ability, background, and repeated observations of real annual earnings (in 1970 dollars) are used to estimate age-earnings relationships.[6] Since earnings represent repeated observation of the 1917–25 cohort group, any exogenous real earnings growth over the period 1943–70 will be confounded in the age variable. The estimated earnings function is of the form[7]

$$(1) \qquad Y(Age_{it}, Sch_i, IQ_i, Soc_i) = \sum_{k=0}^{3} \sum_{j=0}^{2} \sum_{l=0}^{2} \alpha_{kjl} Age_{it}^{k} \cdot Sch_i^{j} \cdot IQ_i^{l} + \sum_{q=1}^{7} \Gamma_q Soc_{qi} + \delta_i + \epsilon_{it}$$

[3]Annual earnings are gross earnings from employment including self-employment. Earnings are not net of taxes. If taxes are roughly proportional then the coefficient of variation and Gini inequality measures will be little affected.

[4]Robert L. Thorndike and Elizabeth Hagen sent a questionnaire to a sample of 17,000 of these men in 1955 which included questions on schooling and 1955 earnings. The *NBER* sent to a subset of these a subsequent questionnaire in 1969 which included additional questions on earnings in later years and questions on schooling and initial job earnings. The data include 5 separate approximately equally spaced points on the age-income profile as well as the year of initial job, year of last full-time schooling, years of schooling, and the twenty separate measures of ability. The age-income points are approximately initial job, 1955, 1960, 1964, and 1968. The observed age range is 19 to 57 years but with less than 1 percent outside the range 19 to 55. The distribution of observations by year are as follows. 3,844 for 1945–52, 1,846 for 1953–57, 3,692 for 1958–62, 1,231 for 1963–66, and 4,774 for 1967–70.

The individuals in the sample differ from the *U.S.* male population as a whole in several ways. First the sample includes a high ability group. All of the men completed high school or high school equivalency examinations, and passed the initial screening for the Air Force flight program. Their general health was better than the general population in 1969. They were more homogenous in height and weight due to military qualifications. They seem to have a high degree to self-confidence and self-reliance. Some of these factors may however be related to the high ability. In addition, all of the men had the G.I. Bill available to help finance their schooling. Other studies based on the *NBER-TH* data include studies by John Hause and Paul Taubman and Terence Wales (1974). Each of these authors considered the effect of ability on earnings in a slightly different way from the one considered here. Thorndike and Hagen give a discription of the early basic data.

[5]The ability index used here is one aggregated by a principal component analysis (the first principal component) from individual indices of reading comprehension, two mathematics tests, two tests of numerical operations, two tests of spatial orientation, speed of identification, and dial and table reading. Since principal components are not scale and origin invariant the aggregate ability index is arbitrarily scaled to have mean one and standard deviation one-quarter.

[6]All 15,387 age-earnings points are simply aggregated, combining the time-series and cross-section aspects.

[7]The earnings function is estimated as that polynomial surface which "best" fits the data in the sense of minimum variance without excessive order. That is, additional order polynomials in age, schooling, and ability are introduced until they fail to reduce error variance significantly at the 5 percent level. The best equation is found to be cubic in *Age*, quadratic in *Sch*, and quadratic in *IQ*. Only *Age* represented a cubic relationship irrespective of the order of entering polynomials. The social variables are entered additively arbitrarily. However, additional investigation suggests that only Jewish religion would significantly interact with age. The earnings function is estimated by ordinary least squares omitting the individual dummy variables representing δ_i. The values of δ_i are estimated in a second stage as the mean residual for each individual. The straightforward dummy variable approach is intractible since 4,699 dummy variables are to be estimated. The reported estimated variance of δ is an unbiased estimate correcting for the finite number of observations per person.

where Age_{it} = age of individual i at observation t; Sch_i = ith individual's years of schooling; IQ_i = ith individual's IQ type ability index; Soc_i is the ith individual's vector of seven social variables including father's and mother's years of schooling, number of siblings, number of childhood family moves, and religion dummy variables for Protestant, Catholic, and Jewish (other religions, no religion, and nonresponses represent the omitted class). The ith individual's permanent deviation from the aggregate earnings function δ_i and the transitory residual ϵ_{it} (itself independently and identically distributed over t and i) are assumed independent of each other and all measured variables. While this earnings function is more complex than the usual linear additive versions, it more completely reveals the interactive nature of the relationships.

Variation in schooling, ability, and background explain 30.1 percent of the variation in annual earnings. The estimated standard deviation of $(\delta_i + \epsilon_{it})$ is $7,997 with a standard deviation of $5,224 and $6,054 for ϵ_{it} and δ_i, respectively.[8] Thus 57 percent of residual variation is explained by individual δ_i differences.[9] This 57 percent may be interpreted as an estimate of the simple correlation between the residuals for any two observations on the same individual. Correspondingly 70 percent of total variation is explained by measured variables plus the permanent component δ_i. These variance components play an important role in human wealth variation as we will see later.

Representative age-earnings profiles based on this earnings function are presented in Figure 1[10] for a Protestant with average levels of other social variables. The earnings profile is shifted vertically by $84 for each additional year of father's education, by $101 for each additional year of mother's education, by −$110 for each additional sibling, by $28 for each childhood family move, and by $345 for Catholic and $4147 for Jewish religion (relative to Protestant).

The life cycle earnings patterns and differences in those patterns due to schooling and ability levels are clearly represented. Earnings rise over the lifetime and rise more rapidly for the more educated and the more able. For example, between the ages 40 and 45, earnings rise at a rate of $556 per year for the college graduate of mean ability while at a rate of $366 for a high school graduate and $880 for a professional or Ph.D. of mean ability. For a college graduate, earnings rise at a rate of $494 per year for an individual one standard deviation below mean ability and $627 for an individual one standard deviation above mean ability.

Both the more educated and the more able initially, prior to age 30, have lower earnings due possibly to higher levels of job training

[8] A more complete specification of the residual structure and alternative estimation procedures is provided in the author (1975) including generalized least squares estimates based on random variance components. All variance estimates presented here are weighted for unequal number of observations for each individual and are corrected for the finite number of multiple observations to make the corresponding estimates unbiased. Similar results from panel data on wider range of birth cohorts observed over a shorter period are reported (i) by the author and Yoram Weiss for a sample of Ph D scientists observed over the decade 1960–70, and (ii) by the author and Robert Willis for a national sample of males from the Michigan Income Dynamics Panel observed over the period 1967–1973

[9] One source of variation in δ is cohort differences within the 1917–25 cohort group due either to cohort differences in, say, schooling quality or to exogenous wage growth occurring between birth cohorts. This source is clearly evident by comparing the mean values of $\hat{\delta}$ across cohorts $1,020 for 1925, 800 for 1924; 18 for 1923; −62 for 1922, −875 for 1921; −745 for 1920; −347 for 1919, −1329 for 1918; and −1,924 for 1917. The variances however do not vary systematically among cohorts

[10] The exact earnings are given by. Y = 4157 + 84 *Father* + 101. *Mother* + 28 *No. Moves* − 110 *No. Sibs* − 295 (*Prot Dummy*) + 50. (*Cath Dummy*) + 3852 (*Jew Dummy*) + 1935 5 (*Sch*) − 296 1 (*Sch Sq*) − 785 9 (*Age*) + 59.4 (*Age Sq*) − 1.085 (*Age Cu*) − 162.6 (*Age*) (*Sch*) + 14 3 (*Age Sq*) (*Sch*) − 23 (*Age Cu*) (*Sch*) + 38 0 (*Age*) (*Sch Sq*) − 2 9 (*Age Sq*) (*Sch Sq*) + .05 (*Age Cu*) (*Sch Sq*) + 2979 1 (*Abil*) (*Age*) − 213 9 (*Abil*) (*Age Sq*) + 3.9 (*Abil*) (*Age Cu*) − 296 2 (*Abil*) (*Age*) (*Sch*) + 21 9 (*Abil*) (*Age Sq*) (*Sch*) − .47 (*Abil*) (*Age Cu*) (*Sch*) − 33 3 (*Abil*) (*Age*) (*Sch Sq*) + 2 4 (*Abil*) (*Age Sq*) (*Sch Sq*) − 032 (*Abil*) (*Age Cu*) (*Sch Sq*) − 2774 5 (*Abil*) (*Sch*) + 459.7 (*Abil*) (*Sch Sq*) − 2106 1 (*Abil Sq*) (*Age*) + 149 6 (*Abil Sq*) (*Age Sq*) − 2 7 (*Abil Sq*) (*Age Cu*) + 364.9 (*Abil Sq*) (*Age*) (*Sch*) − 25.4 (*Abil Sq*) (*Age Sq*) (*Sch*) + 48 (*Abil Sq*) (*Age Cu*) (*Sch*) − 2 9 (*Abil Sq*) (*Age*) (*Sch Sq*) + .17 (*Abil Sq*) (*Age Sq*) (*Sch Sq*) − .008 (*Abil Sq*) (*Age Cu*) (*Sch Sq*) + 463 8 (*Abil Sq*) (*Sch*) − 139.8 (*Abil Sq*) (*Sch Sq*) − 3393 1 (*Abil*) + 3533.0 (*Abil Sq*) *Sch* = years schooling beyond 10, *Age* = age beyond 16

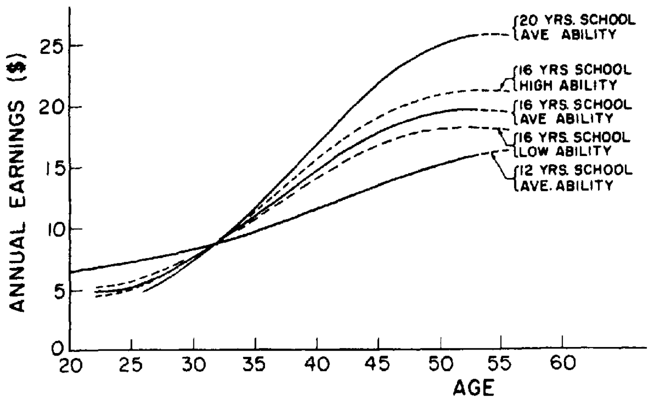


FIGURE 1 ESTIMATED AGE-EARNINGS PROFILES FOR A PROTESTANT WITH AVERAGE VALUES OF OTHER SOCIAL VARIABLES

investment which in turn causes future earnings to rise more rapidly. This empirical relationship illustrates the finding in previous studies (see Zvi Griliches and William Mason and many works cited in Christopher Jencks), that measured cognitive ability has little effect on earnings at early ages It is important to note that almost all studies of the effect of ability on earnings have been based on young men under 35 years of age. Since ability has its greatest effect late in the life cycle, either using young samples or ignoring interaction with age substantially understates the effect of ability.

Another important finding is a strong positive interaction[11] between ability and schooling operating primarily on the age earnings profile. That is, schooling has a greater impact on the age earnings relationship for more able persons, and vice versa. These same positive interactions are also quite evident in their effect on human wealth.

III. Mean Human Wealth

The expected value of human wealth for a given set of measured variables is estimated by summing discounted predicted earnings values, i.e.,

$$(2) \quad MHW(Sch, Abil, Soc) = \sum_{t = Sch+1-10}^{N-16} MY(t, Sch, Abil, Soc)/(1 + r)^t$$

where MY is predicted from the estimated earnings function (1) excluding δ_t, and N is the age of full retirement. The human wealth values presented in Table 1, in 1970 dollars are discounted to age 16 and assume full retirement at age 66.[12] These values correspond to an analysis of the effect of measured variables on the human wealth of a "representative individual."

A striking result is that while schooling has a larger effect on annual earnings at any age than does ability, the effect of schooling on mean human wealth is much more sensitive to discounting than is the effect of ability. When undiscounted, schooling clearly has the dominant effect on lifetime earnings. However, cognitive ability continues to have a positive effect on human wealth at discount rates beyond which the effect of schooling has turned negative. The reason for the difference in sensitivity to dis-

[11]This positive interaction is enhanced by a positive simple coorelation between schooling and ability in the data of 245.

[12]Estimates of human wealth exclude consideration of earnings while in school, for which no data are available, and lower earnings incurred while in military service. Age-earnings profiles are assumed to be flat beyond the upper end of the sample age range, about age 52, since the profiles in Figure 1 appear to peak there. Varying the retirement age between 54 and 70 made no differences in the inequality conclusions.

TABLE 1—MEAN HUMAN WEALTH IN 1970 DOLLARS ASSUMING FULL RETIREMENT AT AGE 66 FOR A PROTESTANT WITH AVERAGE VALUES OF OTHER SOCIAL VARIABLES

Discount Rate	Ability	Years of Schooling: 12	13	14	15	16	17	18	19	20
	Low (.75)	561734	570387	587666	604956	620697	634832	648042	661958	679159
.00	Average (1.00)	590258	590257	604210	624534	649015	677003	708208	742350	780300
	High (1.25)	620864	629259	645082	666390	692553	722883	756962	794537	835340
	Low (.75)	252447	251049	252256	253773	255099	256116	256899	257786	259380
.03	Average (1.00)	260149	255425	255719	258868	264081	270992	279337	288874	299693
	High (1.25)	270434	267564	268909	273106	279615	288035	298072	309510	322146
	Low (.75)	163305	159107	156424	154280	152388	150610	148886	147276	145965
.05	Average (1.00)	166144	160014	156981	155992	156557	158374	161195	164796	169117
	High (1.25)	170960	165399	163198	163166	164790	167737	171754	176643	182230
	Low (.75)	113200	107650	103253	99566	96390	93570	90985	88578	86363
.07	Average (1.00)	113848	107091	102601	99773	98245	97741	98027	98897	100236
	High (1.25)	115908	109217	105424	103441	102776	103129	104270	106020	108228

Note: The ability index is distributed with mean one and standard deviation .25.

TABLE 2—CONTRIBUTION OF SCHOOLING AND ABILITY VARIABLES TO MEAN HUMAN WEALTH
(Discounted at 0–7 Percent)

	0%	3%	5%	7%
College vs. High School				
Low (.75) Ability	58963	2652	−10917	−16810
Average (1.00) Ability	58757	3932	−9587	−15603
High (1.25) Ability	71689	9181	−6170	−13132
Ph.D./Professional vs. College				
Low (.75) Ability	58462	4281	−6423	−10027
Average (1.00) Ability	131285	35612	12560	1991
High (1.25) Ability	142787	42531	17440	5452
Average to Low Ability				
High School	28524	7702	2839	648
College	28318	8982	4169	1855
Ph.D./Professional	101141	40313	23152	13873
High vs. Average Ability				
High School	30606	10285	4816	2060
College	43538	15534	8233	4531
Ph.D./Professional	55040	22453	13113	7992

counting is the period of foregone earnings associated with additional schooling but not with greater ability. An increase in ability for a given schooling level is accompanied by an initial period of slightly lower earnings followed by greater earnings for the remainder of the life cycle; but initial earnings begin at the same age. Additional schooling may be thought of as an investment while additional ability may be thought of as a greater endowment. The effect of a greater endowment of ability is consistent with greater on-the-job training investment which is more than compensated.

Table 2 clearly illustrates the strong positive interaction between ability and schooling in their effect on mean human wealth. More able persons gain more human wealth from additional schooling than do less able persons, and the returns to greater ability are greater at higher levels of schooling. Similarly, the returns to schooling increase rather than decrease with more schooling, and the return to a higher measured ability index is an increasing function of measured ability.[13] For example, discounting at 3 percent, the difference in mean human

[13]This result may be partially due to the composition of the sample studied which includes only highly able and well-educated men

TABLE 3—CONTRIBUTION FOR COLLEGE GRADUATES OF SOCIAL VARIABLES TO MEAN HUMAN WEALTH (Discounted at 0–7 Percent)

Background Variable	0%	3%	5%	7%
Father's Education (one year)	3696	1714	1107	753
Mother's Education (one year)	4444	2061	1331	905
Number Siblings (one more)	−4840	−2245	−1449	−986
Number Pre-High School Moves (one more)	1232	571	369	251
Religion				
Jewish-Protestant	182468	84620	54637	37157
Catholic-Protestant	15180	7040	4545	3091

wealth between a college and a high school graduate is $2,652 at low ability and more than three times that figure ($9,181) at high ability. The corresponding values for Ph.D. versus college are $4,281 and $42,531, respectively

While the set of background data used here is quite limited, we can gain some ideas of their relative importance to human wealth. Mother's education has a slightly larger effect on son's earnings and mean human wealth than does the father's education, by roughly 20 percent. Consider, for example, that these estimates imply that the mother's attainment of a college degree versus a high school degree is associated with an increase of $17,776 in undiscounted lifetime earnings, compared to $14,784 for the same change in father's education. These estimates are roughly 30 percent as large as the effect of the son's own college attainment over high school for an average-ability son. The effect of parents' education is enhanced by their strong positive correlation with each other. The number of siblings has a negative effect on earnings and mean human wealth while the number of pre-high school family moves has an insignificant positive effect. By far the largest background effect is due to religion, particularly if the person is Jewish.[14] See Table 3.

The direct effect of these background variables on earnings and mean human wealth appears to be rather small compared to schooling and ability. These variables also indirectly affect earnings and human wealth through their effect on schooling and ability which is not accounted for here.[15]

IV. Dispersion in Human Wealth

Dispersion in earnings and human wealth arises from the underlying dispersion of their determining characteristics including schooling, measured ability, background, and the unmeasured component δ. One objective here is to move beyond the representative individual notion to estimate the total variation in human wealth A second objective is to assess the relative importance of schooling, ability, and background as determinants of inequality in human wealth Since predicted earnings MY and the estimated value of the individual component δ are orthogonal by construction, the contribution of each to variation in human wealth may be separated and the sum of the two components equals the total

More specifically, the variance in human wealth is estimated in the following manner. First, an estimate of mean human wealth (MHW) is made for each individual in the sample based on his schooling, ability, and background by summing discounted predicted earnings as in equation (2). This estimate corresponds to the expected value of human wealth for a given schooling, ability, and background set. The variance of MHW corresponds to dispersion in human wealth due to measured schooling, ability, and background differences. Next, each individual's human wealth HW is estimated by utilizing the individual's own observed earnings history. The mean discounted residual from MY is calculated so that the present value of each individual's transitory earnings component is zero and thus adds

[14] Part of the effect of this variable may be due to the city size of the respondents' residence since much of the Jewish population resides in the New York metropolitan area which has substantially higher wages than most other parts of the United States.

[15] A detailed study of these indirect effects is presented in the context of a full recursive system relating background, ability, schooling, and lifetime patterns of earnings in the author (1975).

nothing to the variance in human wealth. When the discount rate is zero the resulting mean residual is an unbiased estimate of δ_i. An individual's estimated human wealth *HW* is then *MHW* plus the present value of the mean discounted residual, and the variance in human wealth is the sum of the orthogonal variance components. For calculating variances over individuals each observation is weighted in proportion to the number of age-earnings points observed for that individual. The standard deviations of these discounted mean residuals are \$6,054, \$5,283, \$4,871, and \$4,555 for discount rates of 0, 3, 5, and 7 percent, respectively.

This procedure for calculation of individual human wealth is analogous to estimating the earnings function with discounted earnings values as the dependent variable. Alternative estimates of the variance in human wealth obtained by ignoring the earnings function and estimating each individual's human wealth directly from his observed earnings values were very close to those reported here.

V. Inequality in Earnings vs. Human Wealth

Human wealth *(HW)* is substantially more equally distributed among members of the sample birth cohort than earnings within narrow life cycle age ranges. Inequality in earnings at any stage of the life cycle for men over 30, as measured by either the coefficients of variation or the Gini coefficient[16] is 50 percent larger than inequality in human wealth.[17] This conclusion is not affected by changes in the discount rate.[18]

Since the members of the *NBER-TH* sample are slightly more homogenous than all members of the 1917–25 birth cohort with at least a high school degree, it is useful to compare it with a similar group from the 1960 Census population. The sample cohort group would be ages 35 to 43 in 1960.[19] The corresponding income (from all sources including earnings) inequality among 35-44 year olds with at least a high school degree in the 1960 Census population was .69 by the coefficient of variation and .33 by the Gini coefficient.[20] These differences are not excessively large and are in the expected direction since the *NBER-TH* group is more homogenous.

The difference in inequality between earnings and human wealth is partly but not solely due to compensated differences in lifetime earnings profiles. Inequality in human wealth *HW* is largely dominated by the magnitude of variation is the persistent individual differences. Variation in δ accounts for 40 percent of the total earnings variation, 57 percent of residual earnings variation, and 88 percent of variation in human wealth *HW* (undiscounted).

The importance of this magnitude is illustrated by considering the alternative extreme values of 0 and 100 percent of residual earnings variation due to δ. If all the residual variation were purely random, even within observations for the same individual, then inequality in human wealth would be due solely to the

[16] For detailed discussion of alternative inequality indices see A. B Atkinson.

[17] Inequality among individuals in single year cohorts, rather than the 1917–25 cohort group, as measured by the coefficient of variation are as follows. 39 for 1925; .39 for 1924; 40 for 1923. 44 for 1922; .48 for 1921. 47 for 1920; .53 for 1919; 43 for 1918, and 49 for 1917 The major source of these differences is mean human wealth rather than the standard deviation. The greater mean human wealth for more recent cohorts is due both to differences in mean human wealth caused by schooling, ability and background differences and by differences in mean δ (as noted in fn. 9).

[18] Several retirement ages were considered including a retirement age differing by years of schooling estimated as the mean retirement age for that schooling group based on labor force participation rates. It made virtually no difference in the inequality conclusions reached here

[19] One may be interested in comparing these inequality statements to the more usual cross-section inequality figures Since earnings are roughly uniformly distributed over ages within the sample (with the exclusion of ages over 57), a simple aggregate of the 15387 earnings points over all ages crudely approximates the distribution of earnings of a cross section but with only a narrow cohort observed If there are no cohort or exogenous wage growth effects it is precisely analogous to a cross-sectional earnings distribution Inequality in this aggregate is 75 for the coefficient of variation and .353 for the Gini coefficient.

[20] Calculated from *Census of Population: 1960*, Final Report PC(2)-5A

TABLE 4—DISTRIBUTION OF EARNINGS IN 1970 DOLLARS

Age Group	Mean	Standard Deviation	Coefficient of Variation	Gini Coefficient	Skewness
30–34	10.284	6,115	59	254	6.18
35–39	12,429	7,396	60	281	4 41
40–44	15,110	9,037	60	285	3 18
45–49	18,795	12,260	65	310	3 10

TABLE 5—DISTRIBUTION OF HUMAN WEALTH *HW* IN 1970 DOLLARS ASSUMING FULL RETIREMENT AT AGE 66

Discount Rate	Mean	Standard Deviation	Coefficient of Variation	Gini Coefficient	Skewness
00	674,146	289,380	.43	191	2 69
03	277,533	115,878	42	191	2 94
.05	166,895	69,632	42	186	3 18
07	106,775	45,483	43	.187	3 38

Note: Skewness is measured by the square root of $E(X-\bar{X})^3/S^3$. Coefficient of variation is S/X. Individual observations are weighted by the number of observed age-earnings points.

measured variables,[21] schooling, ability, and background. Under this restriction both the coefficient of variation and Gini coefficient are reduced to one-third the former levels (15 and 7 percent, respectively). Alternatively we may note that schooling, ability, and background account for 10 to 12 percent of the total variation in human wealth, as measured by Var(*MHW*)/Var(*HW*). The remainder is attributed to variation in δ. It is interesting to note at this point that the percent of variation in earnings within the narrow age groups explained by schooling, ability, and background, as measured by Var(*MY/Age*)/Var(*Y/Age*), is also 10 to 12 percent.

At the other extreme all residual differences persist over a lifetime. The upper bound on the coefficient of variation is 50 to 100 percent greater than the predicted values of Table 5. The upper bound also ranges from about the same level to 50 percent larger than the coefficient of variation for earnings within the narrow age groups. See Table 6.

[21]For positive discount rates the corresponding assumption must be that all residual variation is exactly compensated in present value.

TABLE 6—UPPER BOUND ON HUMAN WEALTH *HW* INEQUALITY

Discount Rate (percent)	Mean	Standard Deviation	Coefficient of Variation
0	674,146	401,685	60
3	277,533	181,305	.65
5	166,895	122,331	73
7	106,775	87,603	82

One may reasonably be interested in inequality within schooling or ability groups. The only subgroups with greater human wealth inequality than the aggregate are schooling groups corresponding to some college but not college completion; 13, 14, or 15 years of schooling having a coefficient of variation (*CV*) of 53, 48, and 49, respectively, as compared to the .43 overall, for undiscounted values. This greater inequality is due to greater dispersion, rather than to a lesser mean, relative to other subgroups which are more equally distributed. The greater dispersion is in turn due to greater dispersion in the individual variance component δ, rather than "due to schooling, ability, and background." Across schooling classes, the

CV declines slightly with increased schooling. Across ability groups the CV declines slightly with increased ability. This fall in equality is again due to a less than proportionate rise in dispersion due to δ as the mean human wealth rises with increased ability or schooling. Inequality in annual earnings within schooling and ability groups is at least 50 percent greater than inequality in human wealth within the corresponding subgroups.

VI. A Human Capital Interpretation

The results described above can be interpreted many different ways. The first is simply to accept them as descriptive and interesting in themselves. In the descriptive sense the results require very little structure and each reader is invited to supply his own interpretation of the lifetime earnings patterns, the large individual variance component, and the greater inequality in earnings than in human wealth. Such interpretations might include stories of luck, educational screening, etc.

We would like to supply a structure which is apparently consistent with these results, although admittedly naive; that is, a model of human wealth maximization through investment in human capital. The basic model, first formulated by Ben-Porath, has become a popular vehicle for detailed refinement of models of optimal life cycle investment in human capital. Versions of this model, considered by William Haley, Thomas Johnson, the author, Sherwin Rosen, and T. D. Wallace and Loren Ihnen, assume individual maximization of discounted earnings net of investments. Individuals choose "optimal" schooling periods and lifetime patterns of investment depending on their own endowments, constraints, and abilities. These endogenous decisions then determine a lifetime pattern of earnings with the greatest present value. The individual then maximizes his intertemporal utility subject to this wealth constraint. It is obvious from the formulation of the model that if individuals maximize their human wealth then there will be inequality among individuals in human wealth to the extent that they differ in endownments, constraints, subsidies, and abilities.

While the full model and solution are not presented here,[22] I will outline the formal model and state some relevant predictions. The primary decisions the individual faces are how much human capital to allocate to producing additional human capital (via his personal production function) and how much to spend on purchased inputs at each point of time, and at what age to stop specializing in the production of human capital (i.e., the end of formal schooling). These decisions are influenced by the initial endowment of human capital at the beginning of the planning horizon, the efficiency with which additional human capital can be produced, constraints such as the rate at which human capital deteriorates, the rental rate on human capital, the price of purchased inputs, and the market rate of interest.

The initial endowment of human capital, or earning capacity, has the effect of shifting the earnings profile up or down by the same amount as the change in endowment. It also is inversely related to the optimal length of the schooling period. If we interpret the background variables (all pre-high school except religion) as loosely representing the effect of early public and family investments in children, then their effect will correspond to differences in initial endowment. These variables appear to enter the earnings function linearly as initial earnings capacity would enter if measurable; but this is weak evidence at best

The large individual variance component δ in the earnings function residual is consistent with unmeasured differences in initial earning capacity. It is also consistent with unmeasured differences in investment patterns which are not exactly compensated for in present value. As measured, it includes both of these effects.

One way of interpreting the term ability is differences in the efficiency with which additional human capital can be produced by the

[22] See the author (1975) for details. The results reported here are representative but not identical for all human capital models

individual. Differences in efficiency may represent individual differences in production inputs which are not under the control of the individual, including genetic endowments as well as production inputs provided by society or family. Correspondingly societal and family inputs may be different (and greater) during the formal schooling period than the postschool on-the-job training period. Family and social background variables representing differences in inputs during formal schooling again will shift the earnings function.

Differences in postschooling ability or production efficiency result in earnings profiles which are initially lower for the more able due to a greater level of investment, but rise more rapidly and surpass the earnings of the less able and remain greater throughout the life cycle. The greater investment by the more able is more than compensated in present value. If measured ability (measured just post-high school) represents postschooling production efficiency this prediction is clearly verified by the data as presented in Figure 1. The predicted earnings profile changes just as predicted and human wealth increases with increased ability.

Increased schooling, representing increased investment given initial endowment and ability, increases the period of foregone earnings which is compensated by greater earnings growth and greater earnings late in the life cycles. We thus have the prediction that some earnings inequality is compensated due to differential investment and patterns of returns but that some inequality in human wealth is expected to persist due to differences in endowments, constraints, and abilities. Some of the differences in endowments, constraints, and abilities are represented by measured variables and some are unmeasured and thus captured in the individual residual variance component. The unmeasured individual variance component is the dominant source of the estimated inequality in human wealth indicating a need for further research.

VII. Summary and Conclusions

One of the primary predictions of life cycle models of human capital theory has been a life cycle pattern of investments which decline over time and which yield compensating returns later. Both tend to produce individual earnings profiles which are concave and which rise more rapidly for those with larger early investments. These attributes are roughly confirmed by the data considered here. Both more able and more schooled individuals who are presumably investing more are compensated by more rapidly rising earnings and higher earnings late in the life cycle.

To what extend are these differences in earnings patterns "compensated" in present value? Since each individual is assumed to maximize his lifetime position given his endowments and constraints, there is no presumption that each individual's maximum should be the same; i.e., no inequality in human wealth. We observe substantial variation in human wealth but less inequality in human wealth than earnings within narrow age groups. The coefficient of variation in human wealth is approximately 43 percent compared to 75 percent in earnings, and 60 percent within age groups. The direction of inequality is unambiguous. To the extent that individuals are free to make intertemporal choices about investment in earnings potential, and correspondingly are trading current earnings for later earnings, inequality in current earnings in inappropriate. Inequality in human wealth should be considered.

We observed that the dominant factor in human wealth inequality is the individual variance component representing individual unobserved differences. Only 10 to 12 percent of variation in human wealth is due to variation in measured schooling, ability, and background variables. Roughly the same 10 to 12 percent explanatory power was found within narrow age groups. Whether this is large or small depends on one's point of view given that these are only a few attributes relative to the many which must influence an individual's lifetime.

We also find a positive effect of measured ability on human wealth. While ability has a negligible effect on the earnings of young men, even slightly negative, the effect becomes positive and larger as the men become older.

REFERENCES

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

Z. Griliches and W. M. Mason, "Education, Income and Ability," *J. Polit. Econ.*, May/June 1972, *80*, Part II, S74–S103.

W. J. Haley, "Human Capital: The Choice Between Investment and Income," *Amer. Econ. Rev.*, Dec. 1973, *63*, 929–43.

J. Hause, "Ability and Schooling as Determinants of Lifetime Earnings, or If You're So Smart, Why Aren't You Rich," in F. Thomas Juster, ed , *Education, Income, and Human Behavior*, New York 1975.

H. S. Houthakker, "Education and Income," *Rev. Econ. Statist.*, Feb. 1959, *41*, 24–27.

Christopher Jencks, *Inequality*, New York 1972

T. Johnson, "Returns for Investment in Human Capital," *Amer. Econ. Rev.*, Sept 1970, *60*, 546–60.

L. A. Lillard, "An Essay on Human Wealth," Nat. Bur Econ. Res. monograph, Nov. 1975

———, "The Distribution of Earnings and Human Wealth in a Life Cycle Context," in F. Thomas Juster, ed., *The Distribution of Economic Well-being*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 41, forthcoming.

——— **and Y. Weiss,** "Analysis of Longitudinal Earnings Data: American Scientists 1969–70," Nat. Bur. Econ. Res. working pap. no. 121, Jan. 1976.

——— **and R. Willis,** "Dynamic Aspects of Earnings Mobility." Nat. Bur. Econ. Res. work. pap. no. 150, Sept. 1976.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Review*, Sept. 1975, *65*, 598–609.

S. Rosen, "Income Generating Functions and Capital Accumulation," Harvard Inst. Econ. Res. disc. pap. 306, June 1973.

Paul Taubman and Terence Wales, *Higher Education and Earnings*, New York 1974.

Robert L. Thorndike and Elizabeth Hagen, *Ten Thousand Careers*, New York 1959.

T. D. Wallace and L. A. Ihnen, "Full Time Schooling in Life Cycle Models Human Capital Accumulation," *J Polit. Econ.*, Feb 1975, *83*, 137–55.

B. W. Wilkinson, "Present Values of Lifetime Earnings for Different Occupations," *J. Polit. Econ.*, Dec. 1966, *74*, 556–72.

U.S. Bureau of the Census, *Census of Population:* 1960, Final Report PC(2)-5A, Washington 1963.

Devaluation and Portfolio Balance

By Russell S. Boyer*

The analysis of devaluation has generally been conducted under assumptions which rule out detailed consideration of asset market behavior.[1] Such simplification is justified if there exists only one asset, or if the authorities maintain a "Keynesian-neutral" policy (see S. C. Tsiang) so that the economy behaves as if there is only one asset.[2] Recent analyses have paid careful attention to intertemporal changes in the wealth of the economy (see Robert Mundell, 1971, Eytan Berglas and Assaf Razin, Rudiger Dornbusch, and Anne Krueger), but have continued to assume that wealth takes the form of a single nominal asset denominated in domestic currency.

The purpose of this paper is to introduce into the analysis of devaluation two classes of bonds: nontraded bonds with zero net supply, and traded bonds which can be denominated in either domestic or foreign currency but which are available in perfectly elastic supply. As a result of diversification portfolios are likely to contain an assortment of market instruments, some of which may be denominated in foreign currency. Under these circumstances capital gains or losses become a crucial element in the impact effect of a change in the exchange rate. If, as we shall assume, the devaluation is unanticipated, the quantity of foreign currency denominated assets in domestic portfolios is a parametric initial condition. The discussion is based on the further premise that the central bank maintains fixed exchange rates without sterilization at all times.

The major conclusions of previous research hold even when before devaluation the economy has a position, whether creditor or debtor, in foreign currency denominated market instruments. In the long run, a devaluation causes all nominal variables to rise by the same amount as the exchange rate, whereas all real variables are unchanged. The initial currency-composition of portfolios changes neither this conclusion nor the speed at which the steady state is approached. What it does effect is the initial equilibrium of the economy immediately after the devaluation, and consequently the distance to be traversed to attain the steady state.

As with earlier analyses, this paper shows that a devaluation causes a short-run reduction in real balances, real wealth, and the relative price of the nontraded good.[3] However, the model developed below is able to deal with a number of effects which single nominal asset models fail to capture. In a multiasset model, any exogenous shock causes portfolios to be reshuffled, so that,

*Department of economics, University of Western Ontario, and International Monetary Research Programme, London School of Economics. Financial assistance from the Social Science Research Council, while I was a Research Fellow with the Programme, is gratefully acknowledged. An earlier version of this paper was presented at the Staff Workshop, Department of Economics, University of Southampton, where Ivor Pearce, Gordon Sparks, and A. R. Nobay made helpful comments. This draft has benefited considerably from suggestions by George Borts, June Flanders, Douglas Purvis, Michael Parkin, David Laidler, and an anonymous referee. I of course remain responsible for errors in the analysis.

[1]The role of asset markets is minimized in the "elasticities" approach to devaluation. See Joan Robinson, Fritz Machlup, Gottfried Haberler, Arnold Harberger, Frank Hahn, I. F. Pearce, and Takashi Negishi.

[2]Tsiang points out that many analyses of devaluation have implicit assumptions about neutral behavior on the part of the monetary authorities. He shows that the results of a devaluation depend crucially upon the particular neutral policy that is pursued. If interest rates are kept constant through government action, then Hicks' composite asset theorem permits the treatment of money and bonds as though they are a single asset.

[3]Robinson notes this result of devaluation on real wealth. Sidney Alexander points out the influences that a real balance effect has, but does not emphasize its importance. Recent models containing money and nontraded goods (see Mundell 1971, Berglas and Razin, Dornbusch, and Krueger) all have these results on the relative price and real wealth.

for example, nominal balances become an endogenous variable in the short run. The precise nature of the reshuffling that follows a devaluation depends upon the currency-composition of the initial portfolio. Moreover, yields on assets are endogenous variables in the model in question, and we are able to show that the domestic rate of interest rises in the short run as a result of devaluation.

Section I introduces a model of a small economy which produces and consumes a nontraded good. In Section II this model is linearized by considering differential changes in the values of the endogenous and exogenous variables in the neighborhood of long-run equilibrium, and the short-run response of all endogenous variables to a small devaluation is derived. In Section III, I analyze the consequences of a devaluation for saving and accumulation, thereby investigating the properties of the long-run equilibrium and the process of transition to that equilibrium. Section IV provides a conclusion.

I. The Model

The model is derived from Mundell's (1968) formalization of the Metzlerian framework for an open economy, but incorporates a portfolio balance view of the asset markets. The analysis is focused on a small economy with static expectations which has goods and financial capital mobility with the rest of the world. The smallness of the country means that the foreign currency price of the traded good is given on the world market. With fixed exchange rates and free trade[4] the domestic currency price of the traded good is exogenous also on the domestic market. In the asset markets the economy faces a given rate of return on internationally traded bonds.

In addition there exist nontraded goods and nontraded bonds. Nontraded goods consist of a group of commodities which have such high transport costs that trade in them is not profitable. The nontraded bond, with zero net supply,[5] has some unique features which make it specific to the country which generates it.[6] Thus the price of the nontraded good and the rate of return on the nontraded bond are determined endogenously.

All markets clear instantaneously; in particular, clearing of the labor market assures full employment at all points in time. The supplies of factors are given so that the economy's production point always lies on a given transformation locus between traded and nontraded goods.

The variables whose equilibrium values are determined within the model are: the domestic-currency price index P; the rate of return on the nontraded bond r; real wealth w; the domestic money supply M; the domestic-currency value of internationally traded bonds held domestically B_t; the quantity of foreign currency international bonds in domestic portfolios B; the quantity of local currency international bonds in domestic portfolios B_d; the relative price of the nontraded good p; the current account surplus CAS, and the capital account surplus KAS, both measured in foreign currency; the domestic currency price of the nontraded good P_d; long-run target wealth w^*; and nominal wealth W.

The exogenous variables are: α, the share in

[4]Harry Johnson (1975) recently reemphasized the point that commercial policy can accomplish many of the results of devaluation. It is necessary, as a consequence, to keep the commercial policy stance unchanged during the analysis of a devaluation

[5]The assumption of zero net supply is made because the distribution effects of this asset between the authorities and the private sector may be unimportant or the authorities may maintain a zero net position in it While the use of this particular net supply is analytically convenient, the conclusions do not depend upon this explicitly, but continue to hold so long as this net supply is not substantial (where that term is defined in terms of the partial derivatives of the system) If net supplies of this asset are nonnegligible (but are not so large positively as to change the sign of the system determinant), then two short-run conclusions of this analysis need to be altered With sufficiently large positive net supplies of this asset the interest rate falls on impact due to a devaluation With sufficiently large negative net supplies of this asset real wealth rises

[6]The nontraded bond is meant to represent some aggregate of assets which are poor substitutes for the international bond This would probably include consumer loans, mortgages, equities of local interest, and particularly government bonds with limited marketability In doing empirical work, the appropriate aggregate to use as the nontraded bond on the stock side might be as hard to define as the aggregate of nontraded goods is on the flow side

production of the nontraded good (with $0 < \alpha < 1$), so that $1 - \alpha$ is the share in production of the traded good; e, the domestic-currency price of a unit of foreign currency (the exchange rate defined in the usual way);[7] h, a positive constant that relates the rate of asset accumulation to the discrepancy between desired and actual wealth; and β, the ratio of foreign currency denominated assets to total nominal wealth of the domestic portfolio immediately before devaluation. Units are chosen so that all nominal price variables equal one at the initial equilibrium.

There are six market-clearing equations:

stocks:

(1) money: $P \cdot l(r, w) - M = 0$

$$l_1 < 0,\ l_2 > 0$$

(2) traded bonds: $P \cdot b_t(r, w) - B_t = 0$

$$b_{t_1} < 0,\ b_{t_2} > 0$$

(3) nontraded bonds: $b(r, w) = 0$

$$b_1 > 0,\ b_2 > 0$$

flows:

(4) nontraded goods: $X(p, r, w) = 0$

$$X_1 < 0,\ X_2 < 0,\ X_3 > 0$$

(5) traded goods: $T(p, r, w) + CAS = 0$

$$T_1 > 0,\ T_2 < 0,\ T_3 > 0$$

(6) asset accumulation:

$$h \cdot (w^*(r) - w) + \frac{KAS}{P} \cdot e = 0$$

$$w^{*\prime} > 0$$

In addition, there are six identities defining price and wealth variables:

(7) $p = P_d/e$

(8) $P = P_d^{\alpha} e^{1-\alpha},\ 0 < \alpha < 1$

(9) $W = M + B_t = M + eB + B_d$

(10) $w = W/P$

(11) $\beta = B/W,\ \beta < 1$

The market-clearing conditions, with the usual signs for the partial derivatives, are in the standard format: l is the demand for real balances; b_t is the demand for a real quantity of traded bonds; and b is demand for nontraded bonds. It is assumed that the partial derivatives have the usual signs. These partials are subject to the summing up constraints derived from the definition of wealth, equation (9):

$$l_1 + b_{t_1} + b_1 = 0$$

$$l_2 + b_{t_2} + b_2 = 1$$

These constraints show that one of equations (1)–(3) is redundant. The reason is that with a given level of nominal wealth if two of the three asset markets clear, then the third one clears as well

The net supply of nontraded bonds is zero so they do not appear in the definition of wealth. The ratio of foreign currency denominated assets to total wealth is treated parametrically (equation (11)) because devaluation is unanticipated. Finally, both W and M are assumed to be positive, and β is less than one because the economy is taken to be a creditor in domestic currency assets.

The nontraded goods market clearing condition, where X is the excess demand, is of the usual form. Real income is not included as an argument in any demand function because the model deals only with full-employment real income which can be defined as a constant along a production possibility locus.[8] The equations describing the traded goods market and the asset accumulation process need further discussion. With free trade, the equilibrium current account surplus measured in foreign currency is equal to minus the excess demand for traded goods T, as in equation (5).[9] The excess demand for asset accumulation is equal to desired saving because

[7]The foreign currency price of the traded good is assumed constant, equal to one throughout the analysis

[8]That is, the price deflator is tailored to the production possibility locus so that the real value of income is unchanged as the production point moves around the locus For an analysis of devaluation with underemployment see Tsiang, Ronald Jones, or Tibor Scitovsky

investment is assumed to be zero, in accordance with our assumption of static production. Equation (6) specifies the saving function in a target form: asset accumulation is a constant, h, times the divergence between actual and desired real wealth. The wealth target depends positively on the interest rate. Thus excess demand for asset accumulation is, in equilibrium, equal to the capital account deficit.

These flow market equations are subject to the summing up constraint that all income must be allocated in some fashion. This puts the following conditions on the partial derivatives:

$$X_1 + T_1 = 0$$
$$X_2 + T_2 + hw^{*\prime} = 0$$
$$X_3 + T_3 - h = 0$$

The capital account measured by KAS is defined sufficiently broadly to include all transactions in financial instruments. This definition is used (rather than a further disaggregation of the asset accounts with a portion "above the line" and a portion "below the line") because the analysis here is not sufficiently detailed to deal with the flow demand for any particular asset [10]

Transfers into the economy take the form of interest payments, equal to $r_f B_t/e$ (where r_f is the foreign rate of interest), plus any autonomous transfers T which are taken to be constant in terms of foreign currency. The sum of these two items yields the debt-servicing account surplus:

$$DAS = r_f B_t/e + T \tag{12}$$

These items and transactions in traded goods, the trade account surplus TAS, make up the current account surplus:

$$CAS = TAS + DAS$$

The current account CAS, and the capital account KAS, are defined in an all-inclusive fashion so that they exhaust the transactions recorded in the balance of payments. That is,

$$CAS + KAS = 0 \tag{13}$$

This definition of the balance-of-payments categories is one way of formalizing the flow constraint which the economy faces. To see this, assume that the nontraded goods market clears so that equation (4) holds. If the traded goods market clears (equation (5)) at a value of the current account surplus equal to CAS_0, then the asset accumulation market clears as well, and at a value of KAS equal to $-CAS_0$. Thus, one of equations (4)–(6) is redundant.

II. The Impact Effect of Devaluation

Assume that the economy starts in long-run equilibrium so that the capital account surplus is equal to zero. Wealth accumulation is zero so that markets clear continuously at unchanging prices. These prices can be determined from the general system above. In light of the remarks made in the previous section some of those equations are redundant; furthermore, the model can be solved in a sequential fashion.

The simplest, nonreducible portion of the system may be written as:

$$X\left(P^{1/\alpha}/e^{1/\alpha}, r, \frac{M_0 + B_{d0} + e\beta W_0}{P}\right) = 0$$
$$b\left(r, \frac{M_0 + B_{d0} + e\beta W_0}{P}\right) = 0$$

These equations can be derived by substituting equations (7)–(11) into equations (3) and (4). The exogenous variables are α and the predevaluation composition of portfolios, M_0, B_{d0}, W_0, and β.[11] The exchange rate e is raised exoge-

[9]Albert Hirschman and Richard Cooper have emphasized that one must be specific as to the currency in which any nominal magnitude is measured in a context in which the exchange rate changes

[10]The stock of money (or reserves) is a well-defined function of time. Except for the moment of devaluation this variable has a sensible time derivative The text could be altered to define this time derivative of the stock of money as the portion of the balance of payments accounts below the line without changing any of the conclusions

[11]The subscript 0 denotes the predevaluation level of a variable In a model in which endogenous variables move discontinuously at a number of points in time, this notation serves to indicate the left-hand time limit of a variable The only point in this paper at which the endogenous variables move discontinuously is the moment of devaluation; at all other points there is no difference between the value of an endogenous variable and its time limit This notation therefore serves a dual purpose of indicating the time limit at the moment of devaluation I would like to thank Michael Mussa and Robert Flood for impressing this point upon me

nously in a devaluation.

The exogenous change in the exchange rate affects both the nontraded goods and nontraded bonds markets in ways that can be ascertained from a total differentiation of these equations:

$$(14)\qquad \left(\frac{X_1}{\alpha} - X_2 W\right)dP + X_2 dr + \left(-\frac{X_1}{\alpha} + X_3 \beta W\right)de = 0$$

$$-b_2 W dP + b_1 dr + b_2 \beta W de = 0$$

The determinant of this two-equation system, denoted by Δ, is then equal to

$$\Delta = \frac{X_1 b_1}{\alpha} + (X_2 b_2 - X_3 b_1)W$$

Solving for the changes in the endogenous variables P and r yields

$$(15)\qquad \frac{dP}{de} = \frac{X_1 b_1/\alpha + (X_2 b_2 - X_3 b_1)\beta w}{\Delta}$$

$$\frac{dr}{de} = \frac{X_1 b_2 w(1-\beta)/\alpha}{\Delta}$$

Expression (15) shows that the change in the price index lies within the limits:

$$(16)\qquad \beta \le \frac{dP}{de} \le 1$$

The change in the interest rate on the nontraded bond is bounded by

$$0 \le \frac{dr}{de} \le (1-\beta)\frac{b_2 w}{b_1}$$

These inequalities state that a 1 percent devaluation changes the domestic currency price level by a percentage that is greater than the foreign asset proportion (which can be of either sign) but less than one. This change is not influenced in any important way by the degree of capital substitutability b_1. Interest rates rise by an amount which depends upon the initial currency composition of domestic portfolios and the level of capital substitutability. The higher either of these parameters the smaller the change in interest rates, and in the limit of $\beta = 1$ or $b_1 = \infty$, it is zero.

The change in nominal wealth depends entirely upon the currency composition of portfolios. This can be confirmed by differentiating equation (9); $dW = \beta W_0 de$ or in percentage terms

$$(17)\qquad \frac{1}{W_0}\frac{dW}{de} = \beta$$

Real wealth, on the other hand, falls with a rise in the exchange rate. Real wealth is equal to:

$$w = \frac{W}{P} = \frac{M + e\beta W + B_d}{P}$$

so that

$$\frac{1}{w}\frac{dw}{de} = \beta - \frac{dP}{de} = X_1 b_1(\beta - 1)/(\Delta\alpha)$$

so the limits on the movement of real wealth are

$$(18)\qquad \beta - 1 \le \frac{1}{w}\frac{dw}{de} \le 0$$

If real wealth falls and interest rates rise then real balances must fall as equation (1) indicates. This can be demonstrated most conveniently by differentiating that equation with respect to real wealth and interest rates and substituting the changes in these variables into the derivative. This procedure yields

$$\frac{d\left(\frac{M}{P}\right)}{de} = (1/\Delta)[X_1 w(\beta - 1)(l_2 b_1 - b_2 l_1)/\alpha]$$

which has the same limits as the percentage change in real wealth.[12] Subtracting out the movement of P, the change in the nominal money supply is equal to

$$\frac{1}{M_0}\frac{dM}{de} = [X_1/\alpha(b_1 - (l_1 b_2 - l_2 b_1)w(\beta - 1)) + (X_2 b_2 - X_3 b_1)\beta w] \div \Delta$$

which satisfies the inequalities

[12] This conclusion can be derived from the inequalities $0 \le (l_2 b_1 - b_2 l_1) \le b_1$. The left-hand inequality can be established from the signs of the partials. The right-hand inequality can be proven from the constraints on the sum of the partials in the stock markets.

$$\beta \leq \frac{1}{M_0}\frac{dM}{de} \leq \frac{dP}{de} \leq 1$$

That is, the nominal money supply rises for a devaluing creditor, but may fall for a debtor. In either case, the money supply rises by less on impact than does the price level which, in turn (as shown above), rises by less than the amount of the devaluation.

This complicated expression, with ambiguous sign for $\beta < 0$, can be explained intuitively in terms of the monetary approach to exchange rate changes. A rise in the exchange rate with the domestic-currency price of the nontraded good held constant raises the price index and reduces the real value of wealth. Because the price index effect is greater than the wealth effect, the demand for money rises with a devaluation and the supply of money adjusts endogenously in a fixed exchange rate economy. However, if β is sufficiently large negatively the wealth effect of a rise in the exchange rate dominates the price index effect so that now a devaluation decreases the demand for nominal balances. Over time agents in the economy save and so increase the stock of wealth. The money supply in the long run must be higher than its initial value, although this argument shows that for negative β it may be reduced on impact.

Given the changes in the money supply and nominal wealth brought about by a rise in the exchange rate we may deduce how domestic holdings of international bonds change: $dW/de = dM/de + dB_i/de$ or, from equation (17), $\beta W = dM/de + dB_i/de$. Limits for the change in the money supply set the following bounds on the change in the supply of traded bonds:

$$\beta W - M_0 \leq \frac{dB_i}{de} \leq \beta(W - M_0) \tag{19}$$

The debt servicing account depends on B_t. From equation (13)

$$\frac{dDAS}{de} = r_f\left(\frac{dB_t}{de} - B_t\right)$$

The relationship between initial holdings of money and traded bonds is:

$$B_{t0} = W_0 - M_0$$

Therefore[13]

$$r_f(\beta - 1)W \leq \frac{dDAS}{de} \leq r_f(\beta - 1)(W - M_0) \tag{20}$$

Equation (19) shows that when an economy is indebted in traded bonds, so that $B_t < 0$ (and $W < M_0$), a rise in the exchange rate makes B_t a larger negative number. This implies that the domestic currency value of the debt-servicing outflow (r_f times B_t) is a larger negative number. However, if B_t does not fall by as large a percentage as the exchange rate rises the foreign currency value of this flow will be a smaller negative number, as equation (20) indicates. That is, this portion of the balance-of-payments accounts will improve.

Differentiation of equation (6) shows that a devaluation causes the capital account to worsen. Since that account equals zero initially, the rise in e causes a deficit on capital account:

$$\frac{dKAS}{de} = h(b_1 + w^{*\prime}b_2)X_1w(1 - \beta)/\alpha\Delta \leq 0$$

The reason for the deterioration of the capital account is that interest rates rise, increasing the demand for real wealth at the same time that actual real wealth is reduced. Both of these effects cause an excess domestic demand in the asset accumulation market.

It has been shown elsewhere that $w^{*\prime}/w$ is likely to bear the following relationship to the parameters of the nontraded goods market[14]

$$\frac{w^{*\prime}}{w} \approx -\frac{X_2}{\dfrac{X_1}{\alpha} - X_3w} \tag{21}$$

[13]The expression for the change in the debt-servicing account with a devaluation is:

$$\frac{dDAS}{de} = (\beta - 1)r_f[X_1/\alpha(l_1b_2 + b_1(M_0 - l_2w)) - (X_2b_2 - X_3b_1)W(W - M_0)] \div \Delta$$

[14]See the author. Actually, the *EE* locus in that model is shown to be flatter than the *XX* locus

Under these circumstances

$$h(\beta - 1) \leq \frac{1}{w}\frac{dKAS}{de} \leq 0$$

which is consistent with the change in real wealth, equation (18).

Equation (12) notes that CAS is minus KAS. This implies that the current account improves with a devaluation:[15]

$$0 \leq \frac{1}{w}\frac{dCAS}{de} = -\frac{1}{w}\frac{dKAS}{de} \leq h(1 - \beta)$$

The limits on the movement of the price index in equation (16) can be used to establish the changes in some subsidiary price variables. The domestic currency price of the nontraded good is related to the exchange rate and the price index by equation (8). Its change is given by

$$\frac{dP_d}{de} = \frac{1}{\alpha}\frac{dP}{de} - \frac{1 - \alpha}{\alpha}$$

Substituting in the limits of movement of P we find

$$\frac{\beta - (1 - \alpha)}{\alpha} \leq \frac{dP_d}{de} \leq 1$$

This shows that if the price of the domestic good rises, it rises by less than the increase in the exchange rate. If it falls,[16] the reduction is less than the difference between the ratio of foreign assets to total wealth divided by the proportion of nontraded goods in the price index and the ratio of the proportions of traded to nontraded goods in this index, all multiplied by the change in the exchange rate.

The relative price of the nontraded good in terms of traded goods p, is equal to the ratio of the domestic currency price P_d to the exchange rate e; the change in p is $dp/de = dP_d/de - 1$. Substituting in the limits of movement of P_d establishes $(\beta - 1)/\alpha \leq dp/de \leq 0$. The relative price of the nontraded good deteriorates by less than the difference between the ratio of foreign assets to wealth and one divided by the share of nontraded goods in the price index, all times the rise in the exchange rate.

These results can be compared with the conclusions of previous research which emphasizes the effects of a devaluation on the flow price variables.[17] This comparison demonstrates that the influence of capital gains on foreign currency denominated assets can be incorporated easily into the analysis, independently of the portfolio reshuffling and yield changes which this paper has modeled.

The conclusion of previous research is that a devaluation causes the price index to rise but by not as much as the rise in the exchange rate. In algebraic terms $0 \leq dP/de \leq 1$. This conclusion was derived from the observation that a devaluation with the price level held constant causes the relative price of the nontraded good to fall so that an excess demand for that good arises. If the price level rises with the exchange rate held at its new value then the real value of wealth is reduced, and this causes an offsetting excess supply. Therefore, the price index must rise. However, the price level cannot rise by more than the rise in the exchange rate because this would reverse the relative price movement which originally created the excess demand.

When foreign currency denominated market instruments are introduced, this analysis needs to be modified. The excess demand caused by relative price effects of devaluation with constant price level remains. However, now nominal wealth changes by the percentage βde. Until

[15]In light of the movement of indeterminate sign of DAS, it cannot be demonstrated that TAS is improved with a devaluation. The fact that the debt-servicing account is overwhelmingly likely to worsen makes it even more likely that the trade account improves. However, the stability of the model depends upon the movement of the capital account surplus with a devaluation, as the discussion in Section III demonstrates.

[16]The reason for this possibility is that a rise in the exchange rate can create an excess supply of the nontraded good, requiring a fall in its price to clear that market, even if there are no foreign currency denominated assets. This point is made in the context of business cycle theory by Svend Laursen and Lloyd Metzler. Arnold Harberger (1950) and Johnson (1958) allow for this eventuality in their derivation of the stability condition. Cooper and Dornbusch note this possibility in a more modern framework.

[17]See Mundell (1971), Berglas and Razin, Dornbusch, and Krueger for these results.

the price index moves by that percentage, real wealth for a creditor is increased, reinforcing the excess demand effects of the relative price. In order to eliminate this wealth-induced excess demand, the price level must rise by at least βde. The argument above, that the price index should not rise by the amount of devaluation, continues to hold. This shows that for a creditor the limits on the short-run movement of the price index are $\beta \leq dp/de \leq 1$. The argument is the same when the economy is a debtor in these assets so that this expression is valid also for the case when β is less than zero.

The limits on the movement of this variable in previous research are the same as those derived here when $\beta = 0$. When there are nonzero holdings of foreign assets, whether these limits are narrower or wider depends upon whether the economy is a net creditor or debtor. These altered limits on the change in P carry over to the other price variables.

III. The Transition to Long-Run Equilibrium

The nature of the steady-state equilibrium following a devaluation can be established very quickly. A condition of long-run equilibrium is that the capital account be in balance, so that the quantity of nominal assets is constant over time. Under these circumstances clearing of markets occurs at unchanging prices.

Combining equation (6) the saving function, with KAS equal to zero, with equation (3), clearing of the nontraded bond market, we see that the long-run equilibrium is attained when both the real level of wealth and the rate of interest are restored to their original values. For those values of w and r, the nontraded goods market clears at a relative price of that good equal to its original value. This implies that the domestic currency price of that good must rise as much as the exchange rate does; the price level therefore rises by that percentage. From the money market clearing condition it can be ascertained immediately that the money supply rises by that amount as well. The value of B_f also rises by the percentage of devaluation so that all nominal variables have risen by that percentage and all real variables are restored to their original values. These long run effects are consistent with the conclusions of the "monetary approach" to the analysis of devaluation.[18]

The differential equation which describes the movement to long-run equilibrium is: $dW/dt = -KAS \cdot e$. Taking the derivative of this equation with respect to nominal wealth we obtain

$$\frac{d\left(\frac{dW}{dt}\right)}{dW} = -\frac{dKAS}{dW} \cdot e \tag{22}$$

in the vicinity of long-run equilibrium. This is the speed of adjustment of the economy to its long-run equilibrium, for it shows the rate $d(dW/dt)$ at which any excess nominal balances dW are eliminated from the economy.

This speed of adjustment can be expressed in terms of the partial derivatives above by totally differentiating the system with respect to nominal wealth and solving for $dKAS/dW$. This yields

$$\frac{dKAS}{dW} = h - h\frac{dP}{dW}w - h \cdot w^{*\prime}\frac{dr}{dW}$$

But $$\frac{dP}{dW} = (1/\Delta)(X_2 b_2 - X_3 b_1)$$

and $$\frac{dr}{dW} = -\,b_2 X_1/\alpha/\Delta$$

so that $$\frac{dKAS}{dW} = h(X_1/\alpha)\,(b_1 + b_2 w^{*\prime})/\Delta$$

Using equality (21) this simplifies to

$$0 < \frac{dKAS}{dW} \leq h \tag{23}$$

This expression is positive and shows that an increase in nominal wealth in the steady state causes a capital account surplus. This means that the differential equation (22) describing nominal wealth holdings of the economy is stable with a speed of adjustment, denoted by s, equal to

$$s = h(X_1/\alpha)\,(b_1 + b_2 w^{*\prime})/\Delta$$

[18] See Mundell (1971), Dornbusch, and Johnson (1975) for analyses in the spirit of the "monetary approach."

Expression (23) shows that the economy moves towards its long-run equilibrium at a speed less than h.

The solution of that differential equation is simply

$$W(t) = (W_i - W_f) \exp(-st) + W_f$$

where W_i is the impact value of nominal wealth, W_f is the long-run value of wealth, and t is time measured in units consistent with parameter h and equal to zero at the moment of devaluation. With W_0 equal to the predevaluation level of nominal wealth, W_i is equal to $W_0(1 + \beta de)$, W_f equals $W_0(1 + de)$ so the explicit form of this equation is

$$W(t) = W_0(((\beta - 1)de) \exp(-st) + (1 + de))$$

Furthermore, nominal wealth is the only state variable of this system. Therefore, the time profile of any endogenous variable n is of the form

$$n(t) = (n_i - n_f) \exp(-st) + n_f$$

where n_i is the impact value and n_f is the final value. The price level, for example, follows the pattern

$$P(t) = \left(\frac{(X_2b_2 - X_3b_1)w(\beta - 1)de}{\Delta}\right) \exp(-st) + (1 + de)$$

These results show that the currency composition of portfolios affects neither the steady state nor the speed with which the economy moves to that equilibrium. This is not surprising. the target saving function determines the level of real wealth in the long run independently of the parameter β; and the exchange rate is kept constant after devaluation so there are no subsequent capital gains or losses.

Currency-composition is crucial to determining the change in nominal wealth at devaluation, and consequently the initial equilibrium. A higher value for β causes a larger pure capital gain with a rise in the exchange rate, so that nominal wealth is increased. The discrepancy between the impact value of nominal wealth and its predevaluation level augmented by the percent of devaluation is the saving that the economy must accomplish to get to the steady state.

This argument shows that the path to the long-run equilibrium is the locus of points of short-run equilibrium as the parameter β is altered. For higher values of β the economy is closer to its long-run equilibrium, and the amount that it needs to save is lower. In the limit when $\beta = 1$, the economy moves directly to its long-run equilibrium since the capital gains due to the devaluation are sufficiently large that the rise in the price index has no real effects.

IV. Conclusions

This paper has investigated both impact and long-run effects of a devaluation for an economy which starts in full equilibrium. It has shown that the changes in the price variables in the short run are within narrower or wider limits than are derived when portfolio considerations are not included, depending upon whether the economy is a creditor or debtor in foreign currency denominated assets. In addition, on the flow side the paper has demonstrated that the current account moves into surplus whereas the capital account goes into deficit. The debt-service account moves in an indeterminate fashion although it is very likely to worsen in terms of foreign currency.

A portfolio balance model of the asset markets permits an analysis of the reshuffling of asset holdings due to a devaluation. Interest rates rise at home, and both real wealth and real balances fall. However, the movement of all other variables depends upon whether the economy is a creditor or debtor in foreign currency denominated assets. The results of the "monetary approach" obtain in the long run when all real variables are restored to their original values and all nominal variables increase by the percentage of devaluation.

The general conclusion is that portfolio balance considerations can easily be incorporated into the analysis of devaluation. When the economy is a creditor in foreign currency denominated market instruments the results of a devaluation are more predictable then when these considerations are ignored.

REFERENCES

S. S. Alexander, "Effects of a Devaluation on a Trade Balance," *Int. Monet. Fund Staff Pap.*, Apr. 1952, *2*, 263–78

E. Berglas and A. Razin, "Real Exchange Rate and Devaluation," *J. Int. Econ.*, Feb. 1973, *3*, 179–91.

R. S. Boyer, "Commodity Markets and Bond Markets in a Small, Fixed-Exchange-Rate Economy," *Can. J. Econ.*, Feb. 1975, *8*, 1–23.

Richard N. Cooper, "Currency Devaluation in Developing Countries," *Princeton University International Finance Section Essays in International Finance*, *86*, 1971.

R. Dornbusch, "Devaluation, Money and Non-traded Goods," *Amer Econ. Rev.*, Dec. 1973, *63*, 871–80.

G. Haberler, "The Market for Foreign Exchange and the Stability of the Balance of Payments: A Theoretical Analysis," reprinted in Richard N. Cooper, ed., *International Finance*, Middlesex 1969, ch. 6.

F. H. Hahn, "The Balance of Payments in a Monetary Economy," *Rev. Econ. Stud.*, Feb. 1959, *26*, 110–25.

A. C. Harberger, "Currency Depreciation, Income, and the Balance of Trade," *J. Polit. Econ.*, Feb. 1950, *58*, 47–60.

A. O. Hirschman, "Devaluation and the Trade Balance," *Rev. Econ. Statist.*, Feb. 1949, *31*, 50–53.

H. G. Johnson, "The Transfer Problem and Exchange Stability," *J. Polit. Econ.*, June 1956, *64*, 212–25

———, "The Monetary Approach to the Balance of Payments," International Monetary Research Programme disc. pap., London School of Economics 1975.

R. W. Jones, "Depreciation and the Dampening Effect of Income Changes," *Rev. Econ. Statist.*, Feb. 1960, *42*, 74–80.

A. O. Krueger, "The Role of Home Goods and Money in Exchange Rate Adjustments," in Willy Sellekaerts, ed., *International Trade and Finance: Essays in Honor of Jan Tinbergen*, New York 1974, ch. 7.

S. Laursen and L. A. Metzler, "Flexible Exchange Rates and the Theory of Employment," *Rev. Econ. Statist.*, Nov. 1950, *32*, 281–99.

F. Machlup, "The Theory of Foreign Exchanges," reprinted in Howard S. Ellis and Lloyd A. Metzler, eds, *Readings in the Theory of International Trade*, Philadelphia 1949, ch. 5.

R. A. Mundell, "The International Disequilibrium System," reprinted in his *International Economics*, New York 1968, ch. 15.

———, "Devaluation," in his *Monetary Theory: Inflation, Interest, and Growth in a World Economy*, Pacific Palisades 1971, ch. 9.

T. Negishi, "Approaches to the Analysis of Devaluation," *Int. Econ. Rev.*, June 1968, 9, 218–27.

I. F. Pearce, "The Problem of the Balance of Payments," *Int. Econ. Rev.*, June 1961, *2*, 1–28.

J. Robinson, "The Foreign Exchanges," reprinted in Howard S. Ellis and Lloyd A. Metzler, eds., *Readings in the Theory of International Trade*, Philadelphia 1949, ch. 4.

Tibor Scitovsky, *Money and the Balance of Payments*, New York 1968.

S. C. Tsiang, "The Role of Money in Trade Balance Stability: Synthesis of the Elasticity and Absorption Approaches," reprinted in Richard N Cooper, ed., *International Finance*, Middlesex 1969, ch. 6.

Price Dependent Preferences

By ROBERT A. POLLAK*

Preferences for goods may depend on prices because people judge quality by price or because a higher price enhances the "snob appeal" of a good.[1] Under some circumstances, judging the quality of a product by its price is a rational strategy for an uncertain consumer; presumably, it will be most satisfactory when other consumers are "experts," so that their correct assessments of quality are reflected in the market demand curves facing firms and when our uncertain buyer's tastes (although not his knowledge) coincide with those of the experts. In this paper I suggest a mechanism for incorporating price dependent preferences into demand analysis. The basic trick is to distinguish between "market prices"—the prices which enter the budget constraint—and "normal prices"—the prices which influence preferences. In the final section I discuss briefly the problem of welfare evaluation when preferences depend on prices, but the focus of this paper is on the implications of price dependent preferences for individual demand behavior.

When market prices and normal prices are treated as distinct and independent variables, the resulting model is extremely tractable. I introduce the "market price demand functions" in Section I. They show demand as a function of market prices, total expenditure, and normal prices. Viewed as functions of market prices and total expenditure, these demand functions exhibit all the properties of traditional demand theory.

To complete the normal price model of price dependent preferences it is necessary to specify both the way preferences depend on normal prices and the process by which normal prices are determined. The "normal price function" specifies normal prices as a function of current and past prices. Our casual understanding of both judging quality by price and snob appeal suggests that the price variables which influence tastes are not simply current prices, but some more complex construct related to past as well as current prices. The normal price specification is compatible with this insight, although it is also compatible with the two polar specifications in which normal prices depend exclusively on past prices or exclusively on current prices. The easiest case is the one in which normal prices depend on past prices but not on current prices; in this case the market price demand functions are also short-run demand functions. The other polar case corresponds to the usual "simultaneous specification" of price dependent preferences in which tastes depend on current prices but not on past prices; in this case, the distinction between normal and market prices is only an analytical one, but even here the use of normal prices allows us to distinguish the role of prices as determinants of preferences from their role as determinants of the constraint.

The "relative price hypothesis" postulates that preferences are influenced by relative rather than absolute normal prices. In the two polar cases just described, this implies that both the short- and long-run demand functions are homogeneous of degree zero in current prices and total expenditure (i.e., the absence of "money illusion"). When normal prices depend on both current and past prices, the relative hypothesis does not imply the absence of money

*University of Pennsylvania This research was supported in part by grants from the National Science Foundation I am grateful to Bruce Dieftenbach, Stephen Ross, and an anonymous referee for helpful comments

[1]A third reason why preferences might depend on prices is related to the treatment of money and other financial assets If such assets are included in the utility function, then the prices of goods must also appear As Paul Samuelson says, "The amount of money which is needed depends on the work that is to be done, which in turn depends upon the prices of all goods ..." (p 119) I shall use the phrase "price dependent preferences" to refer to "judging quality by price" and snob appeal, and ignore the dependence which arises in connection with financial assets.

illusion in the short run, but it does imply its absence in the long run.

In Section II, I consider the "normal price demand functions," the demand functions implied by the normal price model when normal and market prices coincide. We can interpret these demand functions as reflecting "long-run" or "steady-state" behavior. If normal prices always coincide with current prices we have the simultaneous specification, and these are the short-run demand functions as well. One can always find a price dependent preference ordering which will generate these normal price demand functions, but a system of normal price demand functions does not correspond to a unique price dependent ordering. I show that the class of price dependent preference orderings which can rationalize a given system of normal price demand functions is very large and argue that this is why a system of demand functions derived from a price dependent preference ordering can sometimes be rationalized by a preference ordering which is independent of prices.

Under the relative price hypothesis, the only restriction other than continuity on the demand functions generated by a simultaneous price dependent preference ordering is that they are homogeneous of degree zero in prices and expenditure and satisfy the budget constraint. Put another way, any system of continuous demand functions homogeneous of degree zero in prices and expenditure which satisfies the budget constraint can be rationalized by a simultaneous price dependent preference ordering satisfying the relative price hypothesis.

In Section III, to conclude, I discuss two alternative interpretations of price dependent preferences, the "conditional" and the "unconditional," and discuss the implications of such preferences for welfare economics. In conditional models of price dependent preferences, the objects of choice are commodity bundles and the preference ordering $R(P)$ depends on prices. In unconditional models, the objects of choice are "quantity-price situations." In the first two sections, I interpret price dependent preferences in terms of the conditional model, but from the standpoint of demand behavior, the two are indistinguishable. For welfare economics, the difference between the two models is significant, since the conditional model does not permit comparisons of situations in which prices differ. However, the unconditional ordering cannot be inferred from observations of the household's market behavior or its behavior in other situations in which prices are fixed and the objects of choice are commodity bundles.

The literature on price dependent preferences is sparse. The snob hypothesis is usually associated with Thorstein Veblen, although the idea itself can be traced back much further. Tibor Scitovsky seems to have been the first to advance the hypothesis of judging quality by price, but his discussion does not focus on the implications of the hypothesis for demand behavior. The basic idea has received little subsequent attention.[2] Harvey Leibenstein draws attention to price dependent preferences without referring to Scitovsky, but does not systematically explore their implications for demand theory. Peter Kalman and M. G. Allingham and Michio Morishima consider the appropriate generalization of the Slutsky equation when prices enter the utility function; their results depend on the unconditional interpretation of the price dependent preference ordering and I will not discuss them further. Roger Alcaly and Alvin Klevorick treat prices as characteristics of goods in a Lancaster-type model. This seems to exhaust the theoretical literature on price dependent preferences.[3]

I. The Normal Price Specification and Market Price Demand Functions

In this section I introduce the "normal price

[2]Price dependent preferences are not mentioned by Hendrik Houthakker in his classic survey of demand theory, nor is the Scitovsky paper listed in the fifty-five page bibliography on preferences, utility, and demand compiled by M. Aoki, John Chipman, and Peter Fishburn in the Minnesota symposium volume edited by Chipman et al.

[3]Little empirical work has been done in this area, a paper by André Gabor and Clive Granger suggests that price dependent preferences are of substantial empirical importance, but their framework is sufficiently far removed from that of traditional economic theory that it is difficult to interpret their results.

specification" of price dependent preferences. The basic idea is to distinguish notationally and conceptually between the two roles which prices play in a model in which they influence tastes. I call the prices which influence preferences normal prices and denote them by P^N; market prices, the prices which enter the budget constraint, are denoted by P^M.

There are two types of demand functions in the normal price specification, the market price (*MP*) and the normal price (*NP*) demand functions. The *MP* demand functions $Q = h(P^M, \mu, P^N)$ are found by maximizing the utility function $V(Q; P^N)$ subject to the budget constraint $\Sigma p_k^M q_k = \mu$, where μ denotes total expenditure.[4] The *NP* demand functions $Q = H(P, \mu)$ show the consumption pattern implied by the *MP* demand functions when normal and market prices coincide: $H(P, \mu) = h(P, \mu, P)$.

It is plausible that normal prices depend on both current and past prices. In the polar case in which normal prices depend exclusively on past prices, the *MP* demand functions are short-run demand functions, and the *NP* demand functions are long-run or steady-state demand functions. If normal prices depend on both current and past prices, the *NP* demand functions are still long-run demand functions, but the *MP* demand functions no longer correspond to short-run demand behavior.

If normal prices depend exclusively on current prices, then normal prices and market prices coincide; this "simultaneous specification" is the only one which has been discussed in the literature on price dependent preferences. In this case the distinction between normal prices and market prices is only an analytical one, but even here the distinction is a useful conceptual device for separating the role of price as a determinant of preferences from its role as a determinant of the feasible set.

The *MP* demand functions, viewed as functions of market prices and expenditure, exhibit all of the properties which traditional theory ascribes to demand functions. They are homogeneous of degree zero in market prices and expenditure, satisfy the budget constraint as an identity, and the implied substitution matrix is symmetric and negative semidefinite.[5] These results depend crucially on holding normal prices fixed and viewing demand as a function of market prices, and they follow immediately from the observation that the *MP* demand functions are derived by maximizing a "well-behaved" utility function $V(Q; P^N)$, subject to a budget constraint $\Sigma p_k^M q_k = \mu$. Since preferences are independent of the prices which enter the budget constraint, the situation is precisely the same as in traditional demand theory; normal prices are simply parameters which shift the utility function and cause no more difficulty than the use of race or age as taste parameters. When the prices which influence preferences are related to the prices which enter the budget constraint, the situation is quite different. The homogeneity of the *MP* demand function does not imply that the short-run demand functions are homogeneous of degree zero in prices and expenditure. With price dependent preferences, changes in current prices alter both tastes and the constraint, but the *MP* demand functions hold tastes fixed and isolate the effect of the change in the constraint.

To examine the way the *MP* demand functions depend on normal prices, it is necessary to specify more precisely the way normal prices influence preferences. I impose two restrictions. The first requires a nonnegative relationship between the normal price of a good and an individual's evaluation of it; the second requires that preferences depend on relative rather than absolute normal prices.[6] Our intuitive understanding of price dependent preferences as reflecting snob appeal or judging quality by price requires a nonnegative relationship between the

[4] I have used the word "expenditure" rather than "income" to mean total expenditure on goods, but I have somewhat inconsistently retained the phrase "income-consumption curve."

[5] It is useful to think of the substitution matrix $S = [s_{ij}]$ in terms of the ordinary rather than compensated demand functions

$$s_{ij} = \frac{\partial h^i}{\partial p_j} + h^j \frac{\partial h^i}{\partial \mu}$$

[6] I also assume that preferences are continuous in Q and P^N.

normal price of a good and the individual's desire for it. The most straightforward formalization of this requirement is in terms of binary choice behavior.[7] We postulate that an increase in the price of the ith good does not cause a commodity bundle containing more of the ith good to decline in the individual's preference ordering relative to one containing less.[8]

The relative price hypothesis postulates that preferences depend on relative rather than absolute normal prices. Technically, I assume that the preference ordering $R(P^N)$ is unaffected by a proportional change in all normal prices. That is, if Q^* is preferred to $\hat{Q}$ at normal prices $\bar{P}^N$, then Q^* is preferred to $\hat{Q}$ at normal prices $\lambda\bar{P}^N$ for all $\lambda > 0$.[9] The relative price hypothesis implies that preferences are independent of the nominal units in which normal prices are measured; that is, there is an absence of money illusion in the process of judging quality by price and in the assessment of the prestige value of goods.[10] Hence, the MP demand functions are unaffected by a proportional increase in all normal prices:

$$h(P^M, \mu, \lambda P^N) = h(P^M, \mu, P^N)$$

for all $\lambda > 0$. I postpone further discussion of the implications of the relative price hypothesis until the end of this section, after I have examined the short-run demand functions.

To specify fully the normal price model we must describe the determination of normal prices as well as the relationship between normal prices and preferences. The normal price function specifies the relationship between normal prices and actual prices. The normal price model of price dependent preferences is somewhat more tractable if normal prices depend only on past prices, but I shall emphasize a formulation in which they depend on both current and past prices. Without specifying the length of the time period, it is difficult to assess the plausibility of various specifications of the normal price function. If the period is very short (consider the continuous time model as a limiting case), the "natural" specification is one in which normal price is independent of current price. If the period is very long (say, a decade or a generation) the simultaneous specification in which normal price is equal to current price is relatively more attractive. I shall focus on the intermediate case in which normal prices depend on both current and past prices.

The normal price function, $N(P_t, P_{t-1}, \ldots)$ specifies normal price as a function of current and past prices: $P_t^N = N(P_t, P_{t-1}, \ldots)$, or to

[7]More specifically, binary choice behavior in situations in which the objects of choice are commodity bundles. That is, an individual is faced with a price vector $\hat{P}$ and then offered a choice between two commodity bundles. The two bundles need not have the same value at prices $\hat{P}$, but the individual is to choose one of them for his own consumption. The essence of price dependent preferences is that an individual may strictly prefer Q^* to $\hat{Q}$ in one price situation and $\hat{Q}$ to Q^* in another. The emphasis on binary choice behavior is especially useful in the context of price dependent preferences, focusing on choice situations in which the household's opportunities are not constrained by a budget enables us to examine the role of prices as determinants of tastes in a situation in which the constraint is independent of prices.

[8]Formally, let R denote the binary relation "at least as good as"; I, "indifference", and S, "strict preference". Since these are all price dependent relations, we write $R(P)$, $I(P)$, and $S(P)$. A commodity bundle Q^* contains more of good i than the commodity bundle $\hat{Q}$ if $q_i^* > \hat{q}_i$; Q^* and $\hat{Q}$ may, of course, differ in their other components. Let $\bar{P}$ and $\bar{\bar{P}}$ be two price vectors which differ only in the price of the ith good and for which $\bar{\bar{p}}_i > \bar{p}_i$. We formalize negative responsiveness in terms of the effect of an increase in p_i on the individual's preference between bundles such as Q^* and $\hat{Q}$. An individual is "nonnegatively responsive" to an increase in the price of the ith good if, for Q^*, $\hat{Q}$, $\bar{P}$, and $\bar{\bar{P}}$,

(i) $Q^* S(\bar{P}) \hat{Q}$ implies $Q^* S(\bar{\bar{P}}) \hat{Q}$, and

(ii) $Q^* I(\bar{P}) \hat{Q}$ implies $Q^* R(\bar{\bar{P}}) \hat{Q}$

[9]The relative price hypothesis can be translated from a hypothesis about preference orderings into an equivalent condition on marginal rates of substitution: the marginal rate of substitution involving any pair of goods is homogeneous of degree zero in normal prices

$$\frac{V_i(Q; \lambda P^N)}{V_j(Q; \lambda P^N)} = \frac{V_i(Q; P^N)}{V_j(Q; P^N)}$$

for all $\lambda > 0$.

[10]The relative price hypothesis has not received much attention in the literature on price dependent preferences. Samuelson (p. 119) discusses it in connection with financial assets in the utility function. Allingham and Morishima make an argument which implies a version of the relative price hypothesis in the context of an unconditional model of price dependent preferences. I return to this in Section III. The recent discussion of the presence or absence of money illusion involving Richard Dusansky and Kalman (1974, 1976), Robert Clower and John Riley, and C. Robert Wichers is specifically concerned with financial assets, but is also related to this point

focus on the normal price of a particular good: $p_{it}^N = N^i(P_t, P_{t-1}, \ldots)$. There are a number of interesting special cases, but instead of pursuing them, I shall concentrate on three general conditions (in addition to continuity) which the normal price function is assumed to satisfy: nonnegativity, homogeneity, and convergence.[11]

Nonnegativity: The normal price of a good is nonnegatively related to its price in every period:

$$\frac{\partial N^i}{\partial p_{i\tau}}(P_t, P_{t-1}, \ldots) \geqq 0 \qquad \tau \leqq t$$

If, *ceteris paribus*, the price of good i in the current or some previous period were higher, its normal price in the current period would not be lower. This leaves open the possibility that normal price is independent of either current price or some or all past prices.

Homogeneity: The normal price function is homogeneous of degree one in past and current prices.

$$N(\lambda P_t, \lambda P_{t-1}, \ldots) = \lambda N(P_t, P_{t-1}, \ldots)$$

for all $\lambda > 0$. That is, if all prices were twice as high in both the current and all past periods, then all normal prices would be twice as high.

Convergence: If actual prices converge to $\hat{P}$ then normal prices will approach $\hat{P}$. Formally, let $\{P_t\}$ denote a sequence of price vectors converging to a limit $\hat{P}$ then

$$\lim_{t\to\infty} N(P_t, P_{t-1}, \ldots) = \hat{P}$$

That is, if prices converge to a particular configuration, then normal prices will also converge to that configuration.

A sequential interpretation of the household's decision process is often convenient. In any period a configurattion of past prices is historically given; these past prices, together with current prices, determine a normal price vector P^N through the normal price function. Corresponding to these normal prices is a preference ordering $R(P^N)$ which satisfies all the assumptions of the traditional theory of consumer behavior, and hence can be represented by a utility function, $V(Q; P^N)$. The *MP* demand functions $h(P^M, \mu, P^N)$ are found by maximizing this utility function subject to the budget constraint $\Sigma p_k^M q_k = \mu$.

The short-run (*SR*) demand functions can be distinguished from the *MP* demand functions now that the normal price function has been described. If normal prices depend wholly or in part on current prices, then the *MP* demand functions do not reflect the complete short-run effect of a change in current prices, because they are based on the assumption that normal prices remain fixed This artifice is useful for understanding the workings of price dependent preferences, but the analytical device must not prevent us from examining the full short-run effect of a price change The *SR* demand functions, $h^*(P_t, \mu_t, P_{t-1}, \ldots)$, are found by maximizing the utility function $V[Q_t; N(P_t, P_{t-1}, \ldots)]$ subject to the budget constraint $\Sigma p_{kt} q_{kt} \leq \mu_t$. A change in current prices affects both the constraint and preferences, unless the normal price function is independent of current prices The *SR* demand functions are related to the *MP* demand functions by

$$h^*(P_t, \mu_t, P_{t-1}, \ldots) = h[P_t, \mu_t, N(P_t, P_{t-1}, \ldots)]$$

Hence, the *SR* demand functions record the complete short-run effect of changes in current price, while the *MP* demand functions record only that portion of the short-run effect caused by the change in the constraint and neglect the portion cased by the change in tastes.

The relative price hypothesis implies that the *MP* demand functions are homogeneous of degree zero in normal prices, but it does not follow that the *SR* demand functions are homogeneous of degree zero in prices and expenditure. The *SR* demand functions may still exhibit money illusion, and inflation which does not affect relative prices may nevertheless alter short-run consumption patterns. For example, suppose normal prices are a weighted average of prices in the current and previous periods

[11]The simplest case is one in which normal prices in period t are equal to actual prices in t − 1 A more appealing assumption is that normal prices are a geometrically weighted average of all past prices, different goods might have different "memory coefficients" indicating that the normal prices of some goods are relatively more sensitive

same for every period. Then a proportionate change in all current prices alters relative normal prices, changes short-run preferences, and affects the household's consumption pattern: the *SR* demand functions are not homogeneous of degree zero in current prices and expenditure. In the long run, however, normal prices adjust to the new prices, and the long-run demand functions (i.e., the *NP* demand functions) are homogeneous of degree zero in prices and expenditure.

If normal prices depend exclusively on past prices, then, even without the relative price hypothesis, the *SR* demand functions are homogeneous of degree zero in current prices and expenditure. If normal prices depend exclusively on current prices (i.e., the simultaneous specification), then the relative price hypothesis implies that the *SR* demand functions are homogeneous of degree zero in current prices and expenditure. Thus, the empirical observation that *SR* demand functions exhibit money illusion contradicts the compound hypothesis that price dependent preferences are influenced solely by current prices and that they satisfy the relative price hypothesis, but such observations are not inconsistent with the relative price hypothesis when normal prices depend in part on past prices and respond gradually to changes in current prices.

II. Normal Price Demand Functions

The *NP* demand functions show the consumption pattern implied by the *MP* demand functions when normal prices and market prices coincide. The *NP* demand functions can be interpreted as long-run or steady state demand functions and in the simultaneous specification they are also the short-run demand functions. In this section I examine the properties of the *NP* demand functions and the possibility of rationalizing them with a price dependent preference ordering. I show that the *NP* demand functions generated by the price dependent preference ordering $R(P^N)$ can be rationalized by $R(P)$, but that they can also be rationalized by a wide class of other price dependent preference orderings. Because of this nonuniqueness, it is possible that some *NP* demand systems generated by price dependent orderings can be rationalized by preference orderings independent or prices; I illustrate this possibility with two examples. Finally, I show that any continuous system of *NP* demand functions homogeneous of degree zero in prices and expenditure and satifying the budget constraint can be rationalized by a price dependent preference ordering satisfying the relative price hypothesis. Hence, price dependent preferences imply no additional restrictions on the *NP* demand functions.

The properties of the *NP* demand functions are easily established. Since the *MP* demand functions are homogeneous of degree zero in market prices and expenditure, and as a consequence of the relative price hypothesis are homogeneous of degree zero in normal prices, the *NP* demand functions are homogeneous of degree zero in prices and expenditure. Since the *MP* demand functions satisfy the budget constraint as an identity, so do the *NP* demand functions.[12] But the implied substitution matrix need not be symmetric or negative semidefinite.[13] Thus, price dependent preferences weaken the presumption that demand functions are downward sloping, and an increase in the price of a good can lead a household to consume only that good. As our examples show, price dependent preferences need not lead to such

[12]If we interpret the *NP* demand functions as long-run or steady-state demand functions, then stability depends on the continuity of the *MP* demand functions in both market prices and normal prices and on the convergence of normal prices to the steady-state configuration of market prices. Continuity of the *MP* demand functions in market prices follows from the usual assumptions about preferences, then continuity in normal prices follows from the assumption that preferences are continuous in Q and P^N. Convergence of normal prices to any steady-state price configuration was postulated as a property of the normal price function. The dynamics of the adjustment process are of little analytical interest since they depend only on exogenous variables, but from the standpoint of applied work, the presence of an adjustment process is itself an attractive feature of the normal price specification.

[13]For example, in the linear expenditure system (see following example (ii)), if

$$b_i = \hat{b}_i + \alpha_i p_i^N / \Sigma \alpha_k p_k^N$$

the implied *NP* demand functions do not yield a symmetric substitution matrix. In the Cobb-Douglas case (the linear expenditure system with the b equal to 0) if

$$a_i = \alpha_i (p_i^N)^2 / \Sigma \alpha_k (p_k^N)^2$$

the substitution matrix is symmetric but not negative semidefinite.

drastic results; indeed, the principal difficulty with price dependent preferences is that they lead to virtually no meaningful restrictions on observable behavior, not that their implications are implausible.

The NP demand functions can always be rationalized by a price dependent preference ordering, even though they are defined as the steady-state values of the MP demand functions rather than derived directly from utility maximization.[14] In particular, if the MP demand functions are derived by maximizing the utility function $V(Q;\ P^N)$ subject to $\Sigma p_k^M q_k = \mu$, then the implied NP demand functions can be derived by maximizing $U(Q;\ P) = V(Q;\ P)$ subject to $\Sigma p_k q_k = \mu$. That is, if $R(P^N)$ denotes the price dependent preference ordering which generates the MP demand functions, then $R(P)$ rationalizes the NP demand functions.

Uniqueness of the preference ordering which generates the NP demand functions is a more subtle question than existence. Surprisingly enough, the price dependent preference ordering which rationalizes the NP demand functions is not unique.[15] Two examples provide a useful introduction to the issues:

(i) *The two-good case:* Suppose there are exactly two goods and consider any price dependent preference ordering satisfying the relative price hypothesis. If the Slutsky sign conditions are satisfied, the NP demand functions generated by this price dependent preference ordering can be rationalized by a preference ordering independent of prices.[16]

(ii) *A price dependent linear expenditure system:* Consider the preference ordering which generates the linear expenditure system; the direct utility function is given by $U(Q) = \Sigma a_k log(q_k - b_k)$. Suppose that the b depend on prices and are given by

$$b_i = \hat{b}_i - \frac{\alpha_i}{p_i} A\Pi(p_k^N)^{\alpha_k} \qquad \sum \alpha_k = 1$$

The implied NP demand functions

$$q_i = \hat{b}_i - \frac{\alpha_i}{p_i} \Sigma p_k \hat{b}_k + \frac{a_i}{p_i}\mu + \frac{(a_i - \alpha_i)}{p_i} A\Pi p_k^{\alpha_k}$$

can be rationalized by a preference ordering which does not depend on prices, namely, the preference ordering corresponding to the indirect utility function $\Psi(P, \mu) = [\mu - f(p)]/g(p)$ where $f(p) = \Sigma\, b_k p_k - A\Pi p_k^{\alpha_k}$ and $g(p) = \Pi p_k^{\alpha_k}$.

Systems of NP demand functions can be rationalized by more than one price dependent preference ordering because most of the information contained in such an ordering is not reflected in the demand functions. If we had complete information on binary choices at various sets of prices, then we could infer the price dependent preference orderings which generated them. But we cannot infer the price dependent preference ordering $R(\bar{P})$ from the NP demand functions because the only observations which are geneated by this preference ordering are those which correspond to the price situation $\bar{P}$. Hence, a perturbation of the preference ordering $R(\bar{P})$ outside the neighborhood of the income-consumption curve corresponding to $\bar{P}$ does not affect demand behavior; such a perturbation would alter the decisions made in

[14] The situation here is not analogous to the case of habit formation. In habit formation models, the short-run demand functions depend on past consumption and the long-run demand functions are defined as their steady-state values. In that case, the long-run demand functions can be rationalized only in certain exceptional cases. See my forthcoming paper for a discussion and references to the literature.

[15] The uniqueness of the utility function which represents a particular price dependent preference ordering is a distinct question. Suppose the utility function $V(Q, P^N)$ represents $R(P^N)$. Any increasing transformation of this utility function yields a utility function representing the same preference ordering. Furthermore, because we are concerned only with preserving the preference ordering over commodity bundles, the transformation may itself depend on normal prices. Thus, the utility $W(Q;\ P^N)$ defined by $W(Q;\ P^N) = F[V(Q;\ P^N),\ P^N]$ where $F_1(V,\ P^N) > 0$ is also a representation of the preference ordering $R(P^N)$

[16] If there are exactly two goods, and if the demand functions are homogeneous of degree zero in (P, μ) and satisfy the budget constraint, then it is easy to verify that they must satisfy the Slutsky symmetry conditions. Hence, under these conditions, the demand functions must satisfy what Leonid Hurwicz calls the "mathematical integrability conditions." If they also satisfy the Slutsky sign conditions, then they satisfy the "economic integrability conditions" as well, and hence can be rationalized by a utility function which is independent of prices.

some binary choice situations, but not in the choice situations corresponding to the budget sets implied by prices $\bar{P}$ for any μ. A "revealed preference" approach fails, because demand behavior in different price-expenditure situations is generated by different preference orderings; the corresponding problem in revealed preference terms is that of inferring a preference ordering from a set of observations all of which correspond to the same set of relative prices.[17]

We say that two price dependent preference orderings are "demand equivalent" (D equivalent) if they generate the same NP demand functions, even if they imply different binary choice behavior and generate different MP demand functions. Our examples show that price dependent preference orderings which correspond to different binary choice behavior may be D equivalent, since both examples are of cases in which a system of demand functions generated by a price dependent preference ordering is rationalized by a preference ordering independent of prices. From a more general viewpoint, both are cases in which a demand system generated by one price dependent preference ordering can be rationalized by another, since a preference ordering independent of prices is a degenerate case of a price dependent preference ordering.

A rich class of D equivalent preference orderings can be constructed from a given preference ordering $R(P)$ by suitably modifying any utility function $V(Q; P)$ which represents $R(P)$ so that the NP demand functions remain unaltered. Since the only portion of the preference ordering which affects the NP demand functions is the portion near the graph $[H(P, \mu), P]$, it is this portion which must remain intact. That is, we require a transformation which carries $V(Q; P)$ into a well-behaved utility function, but leaves marginal rates of substitution on the graph $[H(P, \mu), P]$ unchanged. To construct such a transformation, first choose any differentiable function $S(Q, P)$ which is zero if Q lies on the income consumption curve of P.[18] Then define a new price dependent utility function $W(Q; P)$ by $W(Q; P) = F[V(Q; P), P] + S(Q, P)^2\, \Omega(Q, P)$ where $F_1(V, P) > 0$ and $\Omega(Q, P)$ is any differentiable function of Q and P. If $W(Q; P)$ represents a well-behaved preference ordering, then that preference ordering is D equivalent to $R(P)$.[19] This follows immediately from an examination of the implied marginal rates of substitution.[20]

If we begin with a price dependent preference ordering which rationalizes a system of NP demand functions, then this class of transformations permits us to obtain from it a rich class of D equivalent preference orderings. This is possible because the NP demand functions reflect the properties of the price dependent preference ordering in a narrow region of the quantity-price space. Since the preference ordering outside that region has no effect on the NP demand functions, a system of NP demand functions is consistent with a wide class of price dependent preference orderings which are all identical in the relevant region but differ outside it.

Do any "meaningful theorems" about the NP demand functions follow from the price dependent preferences model? We have already seen that the NP demand functions are continuous and satisfy the budget constraint, and, under the relative price hypothesis, are homogeneous of degree zero in prices and expenditure. It was

[17] If one observes both market prices and normal prices and if they vary independently of one another, then revealed preference reasoning can be used to reconstruct the preference ordering $R(P^N)$ from the MP demand functions $Q = h(P^M, \mu, P^N)$.

[18] An example of a function $S(Q, P)$ having this property is the Euclidian distance from the point Q to the income-consumption curve corresponding to P. That is, $S(Q, P) = \inf \rho(Q, \hat{Q})$ where $\hat{Q}$ ranges over all commodity bundles on the income-consumption curve of P.

[19] Even if the price dependent preference ordering $V(Q, P)$ satisfies the relative price hypothesis, the D equivalent preference ordering $W(Q, P)$ need not do so. This is not surprising, since the relative price hypothesis is equivalent to a strong restriction on the MP demand functions which holds for all values of P^N and P^M, not just those at which $P^N = P^M$.

[20] If $\bar{Q}$ lies on the income-consumption curve of $\bar{P}$, then $S(\bar{Q}, \bar{P}) = 0$ and

$$\frac{W_i(\bar{Q}, \bar{P})}{W_j(\bar{Q}, \bar{P})} = \frac{V_i(\bar{Q}, \bar{P})}{V_j(\bar{Q}, \bar{P})}$$

for all, i, j.

shown by example (see fn. 13) that the implied substitution matrix need not be symmetric or negative semidefinite, but perhaps the *NP* demand functions satisfy some more general set of restrictions. To prove that they do not, I shall show that any system of continuous demand equations satisfying the budget constraint and homogeneous of degree zero in prices and expenditure can be generated by some price dependent preference ordering satisfying the relative price hypothesis

In the case of single valued choice sets and single valued demand systems, a preference ordering is said to rationalize a demand system if, for every budget set, the "best" commodity bundle in the budget set coincides with the commodity bundle demanded. There is no difficulty allowing the demand system to be a correspondence rather than a single valued function, or in admitting choice sets which are not singletons It is, however, useful to distinguish between "weak" and "strong" rationalizations: a preference ordering "strongly rationalizes" a demand system if its choice set coincides with the demand system to be rationalized; it "weakly rationalizes" a demand system if its choice set includes the demand system to be rationalized.[21]

DEFINITION: *A preference ordering R strongly rationalizes a demand system $H(P, \mu)$ if and only if*

$$H(P, \mu) = \{Q: \Sigma p_k q_k \leqq \mu \text{ and } \forall Q^*_{\Sigma p_k q_k^* \leqq \mu} QRQ^*\}$$

DEFINITION: *A preference ordering R weakly rationalizes a demand system $H(P, \mu)$ if and only if*

$$H(P, \mu) \epsilon \{Q: \Sigma p_k q_k \leqq \mu \text{ and } \forall Q^*_{\Sigma p_k q_k^* \leqq \mu} QRQ^*\}$$

To demonstrate that there are no additional restrictions on the *NP* demand functions, I prove two results: 1) any system of demand functions which satisfies the budget constraint can be weakly rationalized by a continuous price dependent preference ordering satisfying the relative price hypothesis; and 2) any system of continuous demand functions which is homogeneous of degree zero in price and expenditure, satisfies the budget constraint, and contains no inferior goods can be strongly rationalized by a continuous fixed coefficient preference ordering satisfying the relative price hypothesis.

I begin with the first proposition. Any system of demand functions which is homogeneous of degree zero in prices and expenditure and satisfies the budget constraint can be weakly rationalized by the price dependent utility function

$$V(Q; P) = \sum_{k=1}^{n} p_k q_k$$

The preference ordering corresponding to this utility function is continuous in Q and P and satisfies the relative price hypothesis. For any price-expenditure situation, the highest attainable curve coincides with the budget line, so the implied demand correspondence includes all points on the budget line. Hence, this utility function weakly rationalizes any system of demand functions which satisfies the budget constraint as an identity, even if the demand functions are not continuous or not homogeneous of degree zero in prices and expenditure. One might argue that this is not a proper rationalization of a demand system, since the demand functions are elements of the demand correspondence implied by the preference ordering but are not identical with it. The definitional issue is fairly clear, and rather than debate whether weak rationalization is an appropriate notion, I now turn to the second assertion.

The second proposition concerns rationalizing a system of demand functions which not only satisfies the budget constraint, but is continuous, homogeneous of degree zero in prices and expenditure, and contains no inferior goods. I shall prove that such a system can be strongly rationalized by a price dependent fixed coefficient utility function

$$V(Q; P) + \min\{q_1, f^2(q_2, P), \ldots, f^n(q_n, P)\}$$

which is continuous and satisfies the relative

[21] Marcel Richter (p. 31) uses "rationalize" to mean what I have called "strongly rationalize."

price hypothesis. The proof constructs the functions $\{f^i(q_i, P)\}$ from the demand system $\{H^i(P, \mu)\}$. I first construct a price dependent preference ordering for all strictly positive price vectors on the unit simplex $\Sigma p_k = 1$, $p_i \geqq 0$, and then use the relative price hypothesis to extend this preference ordering to all strictly positive price vectors.

(1) First solve the demand function for good 1, $H^1(P, \mu)$ for μ as a function of q_1 and P, where P is on the unit simplex. That is define the function $\phi^1(P, q_1)$ by $q_1 = H^1[P, \phi^1(P, q_1)]$.

(2) Next substitute $\phi^1(P, q_1)$ for μ in the demand functions $H^2, \ldots, H^n$ to obtain an expression for the demand for each of these $n-1$ goods as a function of P and q_1. That is, define the functions $\phi^i(P, q_1), i = 2, \ldots, n$, by

$$q_i = \phi^i(P, q_1) = H^i[P, \phi^1(P, q_1)]$$

(3) Next, for each i, $i = 2, \ldots, n$, solve $q_i = \phi^i(P, q_1)$ for q_1 as a function of P and q_i. That is, define the functions $f^i(q_i, P)$, $i = 2, \ldots, n$, by $q_i = \phi^i[P, f^i(P, q_i)]$.

(4) Finally, extend the domain of the functions f^i from the unit simplex to all strictly positive price vectors. Letc $\bar{P}$ be any price vector; then there exists a unique price vector P^* on the unit simplex and a $\lambda > 0$ such that $\bar{P} = \lambda P^*$. Define $f^i(q_i, \bar{P})$ by $f^i(q_i, \bar{P}) = f^i(q_i, P^*)$. By construction, $f^i(q_i, P)$ is homogeneous of degree zone in P.

It is easy to verify that the fixed coefficient utility function

$$V(Q, P) = \min\{q_1, f^2(q_2, P), \ldots, f^n(q_n, P)\}$$

is continuous in Q and P, satisfies the relative price hypothesis, and generates the demand functions $\{H^i(P, \mu)\}$. If the demand system contains inferior goods, it appears possible to generalize this construction to obtain indifference curves with linear segments which meet in obtuse rather than right angles.

III. Conclusion: Alternative Interpretations of Price Dependent Preferences

In this section I summarize the implications of price dependent preferences for demand behavior and discuss an alternative interpretation of price dependent preferences in terms of an "unconditional" rather than a "conditional" preference ordering. This distinction has no significance for demand analysis, but it is crucial for welfare economics.

In the last two sections I have discussed a model in which prices influence preferences because individuals regard them as an index of the snob appeal or of the quality of goods. The model distinguishes between normal prices—the prices which influence preferences—and market prices—the prices which enter the budget constraint, and distinguishes among several types of demand functions. The *MP* demand functions hold normal prices fixed and view demand as a function of market prices and expenditure; they exhibit all the properties of traditional demand theory. The *NP* demand functions show the behavior pattern implied by the *MP* demand functions when normal prices and market prices coincide; they are long-run or steady-state demand functions, and in the simultaneous specification they are also the short-run demand functions. The relative price hypothesis postulates that preferences depend on relative rather than absolute normal prices. Most arguments which make money illusion seem plausible can be interpreted as statements about the gradual adjustment of normal prices in response to changes in actual prices, and are thus not inconsistent with the relative price hypothesis. The *NP* demand functions need not exhibit all of the properties of traditional demand theory, and, indeed, there are no restrictions on these demand functions beyond continuity and the budget constraint, and, under the relative price hypothesis, homogeneity of degree zero in prices and expenditure. Furthermore, any system of *NP* demand functions can be rationalized by a large class of distinct price dependent preference orderings, and it is sometimes possible to rationalize the *NP* demand functions derived from a price dependent preference ordering with a preference ordering which is independent of prices.

In conditional models of price dependent preferences, the objects of choice are commodity bundles, and the preference ordering $R(P)$ depends on prices. In unconditional models, the objects of choice are "quantity-price situations." Let $\tilde{R}$ denote a preference ordering over the quantity-price space, and write $(Q^*, P^*)\tilde{R}(\hat{Q}, \hat{P})$ to indicate that the situation (Q^*, P^*) is at least as good as $(\hat{Q}, \hat{P})$. The conditional model has served as the basis for the analysis developed in Sections I and II. However, when prices are fixed and the objects of choice are commodity bundles, the implications of the two models for binary choice behavior are identical; a fortiori, their implications for demand behavior are identical. The unconditional preference ordering over the $2n$ dimensional space of quantity-price situations could be inferred from a sufficiently rich set of observations on binary choices in which the objects of choice are quantity-price situations. But because the unconditional ordering reflects the individual's preference between situations such as $(\bar{Q}, P^*)$ and $(\bar{Q}, \hat{P})$, in which quantities are identical and prices differ, the unconditional ordering cannot be inferred from observations on binary choice situations in which prices are fixed and the objects of choice are commodity bundles.[22]

Every author who deals with price dependent preferences must at least implicitly select either the conditional or the unconditional approach, but the distinction between the two approaches has not been made explicit in the literature. Kenneth Arrow and F. H. Hahn adopt the conditional model, while Kalman and Allingham and Morishima adopt the unconditional one. The relative price hypothesis can be translated into the unconditional model in two distinct ways. The weaker analogue requires only that the implied conditional price dependent preference ordering $R(P)$ satisfy the relative price hypothesis; this guarantees that binary choice behavior over commodity bundles will be independent of the nominal units in which normal prices are measured. The stronger analogue requires the unconditional preference ordering $\tilde{R}$ to be independent of relative prices; technically, $(Q^*, P^*)\tilde{R}(\hat{Q}, \hat{P})$ implies $(Q^*, \lambda P^*)\tilde{R}(\hat{Q}, \lambda\hat{P})$ for all $\lambda > 0$. Allingham and Morishima argue that to deny this implies "that inflation increases the level of utility, which in view of the arbitrariness of the scale of prices is untenable" (p. 252). Their assertion is intelligible only in the unconditional model, but whether the strong form of the relative price hypothesis holds is a difficult empirical question which cannot be settled by data on situations in which prices are given and the objects of choice are commodity bundles.

Compensated or constant utility market price demand functions can be defined in the conditional model. These are based on a particular set of normal prices which serve to fix the preference ordering, and are found by minimizing the cost (defined in terms of market prices) of attaining a particular indifference curve of the preference map corresponding to the normal prices. The market price compensated demand functions, viewed as functions of market prices, exhibit all of the properties ascribed to compensated demand functions in traditonal demand theory. Normal price compensated demand functions cannot be defined in the conditional model because the model does not permit comparisons of alternative normal price situations.

Compensated demand functions can be defined readily in the unconditional model, since it permits comparisons of alternative quantity-price situations. Kalman uses the unconditional model to argue that a generalized Slutsky equation holds under price dependent preferences, but his compensated demand functions cannot be related to observable behavior when the objects of choice are commodity bundles.

The welfare implications of the price depen-

[22] I have denoted the utility function which represents the conditional preference ordering $R(P)$ by $V(Q;P)$, separating the Q and the P by a semicolon to indicate that the utility function represents a preference ordering over Q which is conditional on P. I denote the utility function in the unconditional model by $U(Q, P)$, where the comma between the Q and the P indicates that the utility functions represent a preference ordering over the $2n$ dimensional space (Q, P). The price dependent transformations of the utility function which were admissible in the conditional model are not permitted in the unconditional model. The unconditional utility function $U(Q, P)$ can only be subjected to transformations which preserve the ordering over quantity-price situations, so the only admissible transformations are of the form $W(Q, P) = F[U(Q, P)]$ where F is an increasing function.

dent preferences model depend crucially on whether one adopts the conditional or the unconditional version. Arrow and Hahn (pp. 129–31) consider the conditional model and conclude that the meaning of Pareto optimality is "somewhat obscure." It is clear that the conditional model does not permit interesting welfare judgements, since virtually all of the situations one would like to compare involve different prices, and, hence, are noncomparable. The unconditional model is the only appropriate framework for comparing situations in which both prices and quantities differ. But an individual's unconditional preference ordering cannot be inferred from either his market behavior or from his binary choice behavior in situations in which prices are fixed and the objects of choice are commodity bundles. It is only revealed by his behavior in situations in which the objects of choice are quantity-price situations.

Traditional welfare economics, by comparing situations in which prices differ, rests on the assumption that the unconditional preference ordering is independent of prices. But as long as our observations are limited to market behavior and to binary choice situations in which prices are fixed and the objects of choice are commodity bundles, this assumption is an article of faith, not a testable hypothesis about observable behavior.

REFERENCES

R. E. Alcaly and A. K. Klevorick, "Judging Quality by Price, Snob Appeal, and the New Consumer Theory," *Zeitschrift fur Nationalekonomie*, July 1970, *30*, 53–64.

M. G. Allingham and M. Morishima, "Veblen Effects and Portfolio Selection," in Michio Morishima, ed., *Theory of Demand: Real and Monetary*, Oxford 1973, 242–70.

M. Aoki, J. S. Chipman, and P. C. Fishburn, "A Selected Bibliography of Works Relating to the Theory of Preferences, Utility and Demand," in John Chipman et al., eds., *Preferences, Utility and Demand*, New York 1971, 437–94.

Kenneth Arrow and F. H. Hahn, *General Competitive Analysis*, San Francisco 1971.

John S. Chipman et al., *Preferences, Utility and Demand: A Minnesota Symposium*, New York 1971.

R. W. Clower and J. G. Riley, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Comment," *Amer. Econ. Rev.*, Mar. 1976, *66*, 184–85.

R. Dusansky and P. J. Kalman, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model," *Amer. Econ. Rev.*, Mar. 1974, *64*, 115–22.

——— **and** ——— "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Reply," *Amer. Econ. Rev.*, Mar. 1976, *66*, 192–95.

A. Gabor and C. W. J. Granger, "Price as an Indicator of Quality: Report on an Enquiry," *Economica*, Feb. 1966, *33*, 43–70.

H. S. Houthakker, "The Present State of Consumption Theory," *Econometrica*, Oct. 1961, *29*, 704–40.

L. Hurwicz, "On the Problem of Integrability of Demand Functions," in John Chipman et al., eds., *Preferences, Utility and Demand*, New York 1971, 174–214.

P. J. Kalman, "Theory of Consumer Behavior When Prices Enter the Utility Function," *Econometrica*, July-Oct. 1968, *36*, 497–510.

H. Leibenstein, "Bandwagon, Snob and Veblen Effects in the Theory of Consumers' Demand," *Quart. J Econ.*, May 1950, *64*, 183–207.

R. A. Pollak, "Habit Formation and Long-Run Utility Functions," *J. Econ. Theory*, forthcoming.

M. K. Richter, "Rational Choice," in John Chipman et al., eds., *Preferences, Utility and Demand*, New York 1971, 29–58.

Paul A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

T. Scitovsky, "Some Consequences of the Habit of Judging Quality by Price," *Rev. Econ. Stud.*, 1945, No. 2, *12*, 1–105.

Thorstein Veblen, *The Theory of the Leisure Class*, New York 1899.

C. R. Wichers, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Comment," *Amer. Econ. Rev.*, Mar. 1976, *66*, 186–91.

De Gustibus Non Est Disputandum

By GEORGE J. STIGLER AND GARY S. BECKER*

The venerable admonition not to quarrel over tastes is commonly interpreted as advice to terminate a dispute when it has been resolved into a difference of tastes, presumably because there is no further room for rational persuasion. Tastes are the unchallengeable axioms of a man's behavior: he may properly (usefully) be criticized for inefficiency in satisfying his desires, but the desires themselves are *data*. Deplorable tastes —say, for arson—may be countered by coercive and punitive action, but these deplorable tastes, at least when held by an adult, are not capable of being changed by persuasion.

Our title seems to us to be capable of another and preferable interpretation: that tastes neither change capriciously nor differ importantly between people. On this interpretation one does not argue over tastes for the same reason that one does not argue over the Rocky Mountains—both are there, will be there next year, too, and are the same to all men.

The difference between these two viewpoints of tastes is fundamental. On the traditional view, an explanation of economic phenomena that reaches a difference in tastes between people or times is the terminus of the argument: the problem is abandoned *at this point* to whoever studies and explains tastes (psychologists? anthropologists? phrenologists? sociobiologists?). On our preferred interpretation, one never reaches this impasse: the economist continues to search for differences in prices or incomes to explain any differences or changes in behavior.

The choice between these two views of the role of tastes in economic theory must ultimately be made on the basis of their comparative analytical productivities. On the conventional view of inscrutable, often capricious tastes, one drops the discussion as soon as the behavior of tastes becomes important—and turns his energies to other problems. On our view, one searches, often long and frustratingly, for the subtle forms that prices and incomes take in explaining differences among men and periods. If the latter approach yields more useful results, it is the proper choice. The establishment of the proposition that one may usefully treat tastes as stable over time and similar among people is the central task of this essay.

The ambitiousness of our agenda deserves emphasis: we are proposing the hypothesis that widespread and/or persistent human behavior can be explained by a generalized calculus of utility-maximizing behavior, without introducing the qualification "tastes remaining the same." It is a thesis that does not permit of direct proof because it is an assertion about the world, not a proposition in logic. Moreover, it is possible almost at random to throw up examples of phenomena that presently defy explanation by this hypothesis: Why do we have inflation? Why are there few Jews in farming?[1] Why are societies with polygynous families so rare in the modern era? Why aren't blood banks responsible for the quality of their product? If we could answer these questions to your satisfaction, you would quickly produce a dozen more.

What we assert is not that we are clever enough to make illuminating applications of utility-maximizing theory to all important phenomena—not even our entire generation of economists is clever enough to do that. Rather, we assert that this traditional approach of the

*University of Chicago. We have had helpful comments from Michael Bozdarich, Gilbert Ghez, James Heckman, Peter Pashigian, Sam Peltzman, Donald Wittman, and participants in the Workshop on Industrial Organization

[1]Our lamented friend Reuben Kessel offered an attractive explanation: since Jews have been persecuted so often and forced to flee to other countries, they have not invested in immobile land, but in mobile human capital—business skills, education, etc.—that would automatically go with them. Of course, someone might counter with the more basic query: but why are they Jews, and not Christians or Moslems?

economist offers guidance in tackling these problems—and that no other approach of remotely comparable generality and power is available.

To support our thesis we could offer samples of phenomena we believe to be usefully explained on the assumption of stable, well-behaved preference functions. Ultimately, this is indeed the only persuasive method of supporting the assumption, and it is legitimate to cite in support all of the existing corpus of successful economic theory. Here we shall undertake to give this proof by accomplishment a special and limited interpretation. We take categories of behavior commonly held to demonstrate changes in tastes or to be explicable only in terms of such changes, and show both that they are reconcilable with our assumption of stable preferences and that the reformulation is illuminating.

I. The New Theory of Consumer Choice

The power of stable preferences and utility maximization in explaining a wide range of behavior has been significantly enhanced by a recent reformulation of consumer theory [2] This reformulation transforms the family from a passive maximizer of the utility from market purchases into an active maximizer also engaged in extensive production and investment activities. In the traditional theory, households maximize a utility function of the goods and services bought in the marketplace, whereas in the reformulation they maximize a utility function of objects of choice, called commodities, that they produce with market goods, their own time, their skills, training and other human capital, and other inputs. Stated formally, a household seeks to maximize

$$(1) \qquad U = U(Z_1, \ldots Z_m)$$

with

$$(2) \qquad Z_i = f_i(X_{1i}, \ldots X_{ki}, t_{1i}, \ldots t_{li}, S_1, \ldots S_l, Y_i), \; i = 1 \ldots m$$

where Z_i are the commodity objects of choice entering the utility function, f_i is the production function for the ith commodity, X_{ji} is the quantity of the jth market good or service used in the production of the ith commodity, t_{ji} is the jth person's own time input, S_j the jth person's human capital, and Y_i represents all other inputs.

The Z_i have no market prices since they are not purchased or sold, but do have "shadow" prices determined by their costs of production. If f_i were homogeneous of the first degree in the X_{ji} and t_{ji}, marginal and average costs would be the same and the shadow price of Z_i would be

$$(3) \qquad \pi_i = \sum_{j=1}^{k} \alpha_{ji}\left(\frac{p}{w_1}, \frac{w}{w_1}, S, Y_i\right) p_j + \sum_{j=1}^{l} \beta_{ji}\left(\frac{p}{w_1}, \frac{w}{w_1}, S, Y_i\right) w_j$$

where p_j is the cost of X_j, w_j is the cost of t_j, and α_{ji} and β_{ji} are input-output coefficients that depend on the (relative) set of p and w, S, and Y_i. The numerous and varied determinants of these shadow prices give concrete expression to our earlier statement about the subtle forms that prices take in explaining differences among men and periods.

The real income of a household does not simply equal its money income deflated by an index of the prices of market goods, but equals its full income (which includes the value of "time" to the household)[3] deflated by an index of the prices, π_i, of the produced commodities. Since full income and commodity prices depend on a variety of factors, incomes also take subtle forms. Our task in this paper is to spell out some of the forms prices and full income take.

II. Stability of Tastes and "Addiction"

Tastes are frequently said to change as a result of consuming certain "addictive" goods. For example, smoking of cigarettes, drinking of alcohol, injection of heroin, or close contact with some persons over an appreciable period of

[2]An exposition of this reformulation can be found in Robert Michael and Becker This exposition emphasizes the capacity of the reformulation to generate many implications about behavior that are consistent with stable tastes.

[3]Full income is the maximum money income that a household could achieve by an appropriate allocation of its time and other resources

time, often increases the desire (creates a craving) for these goods or persons, and thereby cause their consumption to grow over time. In utility language, their marginal utility is said to rise over time because tastes shift in their favor. This argument has been clearly stated by Alfred Marshall when discussing the taste for "good" music:

> There is however an implicit condition in this law [of diminishing marginal utility] which should be made clear. It is that we do not suppose time to be allowed for any alteration in the character or tastes of the man himself. It is therefore no exception to the law that the more good music a man hears, the stronger is his taste for it likely to become . . . [p. 94]

We believe that the phenomenon Marshall is trying to explain, namely that exposure to good music increases the subsequent demand for good music (for some persons!), can be explained with some gain in insight by assuming constant tastes, whereas to assume a change in tastes has been an unilluminating "explanation." The essence of our explanation lies in the accumulation of what might be termed "consumption capital" by the consumer, and we distinguish "beneficial" addiction like Marshall's good music from "harmful" addiction like heroin.

Consider first beneficial addiction, and an unchanging utility function that depends on two produced commodities:

$$U = U(M, Z) \tag{4}$$

where M measures the amount of music "appreciation" produced and consumed, and Z the production and consumption of other commodities. Music appreciation is produced by a function that depends on the time allocated to music (t_m), and the training and other human capital conducive to music appreciation (S_m) (other inputs are ignored):

$$M = M_m(t_m, S_m) \tag{5}$$

We assume that

$$\frac{\partial M_m}{\partial t_m} > 0, \ \frac{\partial M_m}{\partial S_m} > 0$$

and also that

$$\frac{\partial^2 M_m}{\partial t_m \partial S_m} > 0$$

An increase in this music capital increases the productivity of time spent listening to or devoted in other ways to music.

In order to analyze the consequences for its consumption of "the more good music a man hears," the production and consumption of music appreciation has to be dated. The amount of appreciation produced at any moment j, M_j, would depend on the time allocated to music and the music human capital at j: t_{m_j} and S_{m_j}, respectively. The latter in turn is produced partly through "on-the-job" training or "learning by doing" by accumulating the effects of earlier music appreciation:

$$S_{m_j} = h(M_{j-1}, M_{j-2}, \ldots, E_j) \tag{6}$$

By definition, the addiction is beneficial if

$$\frac{\partial S_{m_j}}{\partial M_{j-v}} > 0, \text{ all } v \text{ in (6)}$$

The term E_j measures the effect of education and other human capital on music appreciation skill, where

$$\frac{\partial S_{m_j}}{\partial E_j} > 0$$

and probably

$$\frac{\partial^2 S_{m_j}}{\partial M_{j-v} \partial E_j} > 0$$

We assume for simplicity a utility function that is a discounted sum of functions like the one in equation (4), where the M and Z commodities are dated, and the discount rate determined by time preference.[4] The optimal allocation of consumption is determined from the equality between the ratio of their marginal utilities and the ratio of their shadow prices:

$$\frac{MU_{m_j}}{MU_{z_j}} = \frac{\partial U}{\partial M_j} \Big/ \frac{\partial U}{\partial Z_j} = \frac{\pi_{m_j}}{\pi_{z_j}} \tag{7}$$

The shadow price equals the marginal cost of adding a unit of commodity output. The marginal cost is complicated for music appreciation M by the positive effect on subsequent music human capital of the production of music

[4]A consistent application of the assumption of stable preferences implies that the discount rate is zero; that is, the absence of time preference (see the brief discussion in Section VI.)

appreciation at any moment j. This effect on subsequent capital is an investment return from producing appreciation at j that reduces the cost of production at j. It can be shown that the marginal cost at j equals[5]

$$(8) \qquad \pi_{m_j} = \frac{w\partial t_{m_j}}{\partial M_j} - w \sum_{i=1}^{n-j} \frac{\partial M_{j+i}}{\partial S_{m_{j+i}}} \Bigg/ \frac{\partial M_{j+i}}{\partial t_{m_{j+i}}} \cdot \frac{dS_{m_{j+i}}}{dM_j} \cdot \frac{1}{(i+r)^i}$$

$$= \frac{w\partial t_{m_j}}{\partial M_j} - A_j = \frac{w}{MP_{t_{m_j}}} - A_j$$

where w is the wage rate (assumed to be the same at all ages), r the interest rate, n the length of life, and A_j the effect of addiction, measures the value of the saving in future time inputs from the effect of the production of M in j on subsequent music capital.

With no addiction, $A_j = 0$ and equation (8) reduces to the familiar marginal cost formula. Moreover, A_j is positive as long as music is beneficially addictive, and tends to decline as j increases, approaching zero as j approaches n. The term w/MP_{t_m} declines with age for a given time input as long as music capital grows with age. The term A_j may not change so much with age at young ages because the percentage decline in the number of remaining years is small at these ages. Therefore, π_m would tend to decline with age at young ages because the effect on the marginal product of the time input would tend to dominate the effect on A. Although π_m might not always decline at other ages, for the present we assume that π_m declines continuously with age.

If π_z does not depend on age, the relative price of music appreciation would decline with age; then by equation (7), the relative consumption of music appreciation would rise with age. On this interpretation, the (relative) consumption of music appreciation rises with exposure not because tastes shift in favor of music, but because its shadow price falls as skill and experience in the appreciation of music are acquired with exposure.

An alternative way to state the same analysis is that the marginal utility of time allocated to music is increased by an increase in the stock of music capital.[6] Then the consumption of music appreciation could be said to rise with exposure because the marginal utility of the time spent on music rose with exposure, even though tastes were unchanged.

The effect of exposure on the accumulation of music capital might well depend on the level of education and other human capital, as indicated by equation (6). This would explain why educated persons consume more "good" music (i.e., music that educated people like!) than

[5]The utility function

$$V = \sum_{j=1}^{n} a^j U(M_j, Z_j)$$

is maximized subject to the constraints

$$M_j = M(t_{m_j}, S_{m_j}),\ Z_j = Z(x_j, t_{z_j})$$

$$S_{m_j} = h(M_{j-1}, M_{j-2}, \quad , E_j)$$

$$\sum \frac{px_j}{(1+r)^j} = \sum \frac{wt_{wj} + b_j}{(i+r)^j}$$

and $t_{w_j} + t_{m_j} + t_{z_j} = t$,

where t_{w_j} is hours worked in the jth period, and b_j is property income in that period. By substitution one derives the full wealth constraint.

$$\sum \frac{px_j + w(t_{m_j} + t_{z_j})}{(1+r)^j} = \sum \frac{wt + b_j}{(1+r)^j} = W$$

Maximization of V with respect to M_j and Z_j subject to the production functions and the full wealth constraint gives the first-order conditions

$$a^j \frac{\partial U}{\partial Z_j} = \frac{\lambda}{(1+r)^j}\left(\frac{pdx_j}{dZ_j} + \frac{wdt_{z_j}}{dZ_j}\right) = \frac{\lambda}{(1+r)^j}\pi_{z_j}$$

$$a^j \frac{\partial U}{\partial M_j} = \frac{\lambda}{(1+r)^j} \cdot \left(\frac{w\partial t_{m_j}}{\partial M_j} + \sum_{i=1}^{n-j} \frac{wdt_{m_{j+i}}}{dM_j} \cdot \frac{1}{(1+r)^i}\right)$$

$$= \frac{\lambda}{(1+r)^j}\pi_{m_j}$$

Since, however,

$$\frac{dM_{j+i}}{dM_j} = 0 = \frac{\partial M_{j+i}}{\partial S_{m_{j+i}}}\frac{dS_{m_{j+i}}}{dM_j} + \frac{\partial M_{j+i}}{\partial t_{m_{j+i}}}\frac{dt_{m_{j+i}}}{dM_j}$$

then

$$\frac{dt_{m_{j+1}}}{dM_j} = -\frac{\partial M_{j+i}}{\partial S_{m_{j+i}}} \Bigg/ \frac{\partial M_{j+i}}{\partial t_{m_{j+i}}} \cdot \frac{dS_{m_{j+i}}}{dM_j}$$

By substitution into the definition of π_{m_j}, equation (8) follows immediately

[6]The marginal utility of time allocated to music at j includes the utility from the increase in the future stock of music capital that results from an increase in the time allocated at j. An argument similar to the one developed for the price of music appreciation shows that the marginal utility of time would tend to rise with age, at least at younger ages.

other persons do.

Addiction lowers the price of music appreciation at younger ages without any comparable effect on the productivity of the time spent on music at these ages. Therefore, addiction would increase the time spent on music at younger ages: some of the time would be considered an investment that increases future music capital. Although the price of music tends to fall with age, and the consumption of music tends to rise, the time spent on music need not rise with age because the growth in music capital means that the consumption of music could rise even when the time spent fell with age. The time spent would be more likely to rise, the more elastic the demand curve for music appreciation. We can express this result in a form that will strike many readers as surprising; namely, that the time (or other inputs) spent on music appreciation is more likely to be addictive—that is, to rise with exposure to music—the more, not less, elastic is the demand curve for music appreciation.

The stock of music capital might fall and the price of music appreciation rise at older ages because the incentive to invest in future capital would decline as the number of remaining years declined, whereas the investment required simply to maintain the capital stock intact would increase as the stock increased. If the price rose, the time spent on music would fall if the demand curve for music were elastic. Consequently, our analysis indicates that the observed addiction to music may be stronger at younger than at older ages.

These results for music also apply to other commodities that are beneficially addictive. Their prices fall at younger ages and their consumption rises because consumption capital is accumulated with exposure and age. The time and goods used to produce an addictive commodity need not rise with exposure, even though consumption of the commodity does; they are more likely to rise with exposure, the more elastic is the demand curve for the commodity. Even if they rose at younger ages, they might decline eventually as the stock of consumption capital fell at older ages.

Using the same arguments developed for beneficial addiction, we can show that all the results are reversed for harmful addiction,[7] which is defined by a negative sign of the derivatives in equation (6):

$$\frac{\partial S_j}{\partial H_{j-v}} < 0, \text{ all } v \text{ in (6)} \tag{9}$$

where H is a harmfully addictive commodity. An increase in consumption at any age reduces the stock of consumption capital available subsequently, and this raises the shadow price at all ages.[8] The shadow price would rise with age and exposure, at least at younger ages, which would induce consumption to fall with age and exposure. The inputs of goods and time need not fall with exposure, however, because consumption capital falls with exposure; indeed, the inputs are likely to rise with exposure if the commodity's demand curve were inelastic.

To illustrate these conclusions, consider the commodity "euphoria" produced with input of heroin (or alcohol or amphetamines.) An increase in the consumption of current euphoria raises the cost of producing euphoria in the future by reducing the future stock of "euphoric capital." The effect of exposure to euphoria on the cost of producing future euphoria reduces the consumption of euphoria as exposure continues. If the demand curve for euphoria were sufficiently inelastic, however, the use of heroin would grow with exposure at the same time that euphoria fell.

Note that the amount of heroin used at younger ages would be reduced because of the negative effect on later euphoric capital. Indeed, no heroin at all might be used only because the harmfully addictive effects are anticipated, and discourage any use. Note further that if heroin

[7]In some ways, our analysis of beneficial and harmful addiction is a special case of the analysis of beneficial and detrimental joint production in Michael Grossman

[8]Instead of equation (8), one has

$$\pi_{h_j} = \frac{w}{MP_{t_j}} + A_j$$

where $A_j \geq 0$

were used even though the subsequent adverse consequences were accurately anticipated, the utility of the user would be greater than it would be if he were prevented from using heroin. Of course, his utility would be still greater if technologies developed (methadone?) to reduce the harmfully addictive effects of euphoria.[9]

Most interestingly, note that the use of heroin would grow with exposure at the same time that the amount of euphoria fell, if the demand curve for euphoria and thus for heroin were sufficiently inelastic. That is, addiction to heroin—a growth in use with exposure—is the *result* of an inelastic demand for heroin, *not*, as commonly argued, the *cause* of an inelastic demand. In the same way, listening to music or playing tennis would be addictive if the demand curves for music or tennis appreciation were sufficiently elastic; the addiction again is the result, not the cause, of the particular elasticity. Put differently, if addiction were surmised (partly because the input of goods or time rose with age), but if it were not clear whether the addiction were harmful or beneficial, the elasticity of demand could be used to distinguish between them: a high elasticity suggests beneficial and a low elasticity suggests harmful addiction.[10]

We do not have to assume that exposure to euphoria changes tastes in order to understand why the use of heroin grows with exposure, or why the amount used is insensitive to changes in its price. Even with constant tastes, the amount used would grow with exposure, and heroin is addictive precisely *because* of the insensitivity to price changes.

An exogenous rise in the price of addictive goods or time, perhaps due to an excise tax, such as the tax on cigarettes and alcohol, or to restrictions on their sale, such as the imprisonment of dealers in heroin, would have a relatively small effect on their use by addicts if these are harmfully addictive goods, and a relatively large effect if they are beneficially addictive. That is, excise taxes and imprisonment mainly transfer resources away from addicts if the goods are harmfully addictive, and mainly reduce the consumption of addicts if the goods are beneficially addictive.

The extension of the capital concept to investment in the capacity to consume more efficiently has numerous other potential applications. For example, there is a fertile field in consumption capital for the application of the theory of division of labor among family members.

III. Stability of Tastes and Custom and Tradition

A "traditional" qualification to the scope of economic theory is the alleged powerful hold over human behavior of custom and tradition. An excellent statement in the context of the behavior of rulers is that of John Stuart Mill:

> It is not true that the actions even of average rulers are wholly, or anything approaching to wholly, determined by their personal interest, or even by their own opinion of their personal interest. . . . I insist only on what is true of all rulers, viz., that the character and course of their actions is largely influenced (independently of personal calculations) by the habitual sentiments and feelings, the general modes of thinking and acting, which prevail throughout the community of which they are members; as well as by the feelings, habits, and modes of thought which characterize the particular class in that community to which they themselves belong. . . . They are also much influenced by the maxims and traditions which have descended to them from other rulers, their predecessors; which maxims and traditions have been known to retain an ascendancy during long periods, even

[9]That is, if new technology reduced and perhaps even changed the sign of the derivatives in equation (9). We should state explicitly, to avoid any misunderstanding, that "harmful" means only that the derivatives in (9) are negative, and not that the addiction harms others, nor, as we have just indicated, that it is unwise for addicts to consume such commodities.

[10]The elasticity of demand can be estimated from the effects of changes in the prices of inputs. For example, if a commodity's production function were homogeneous of degree one, and if all its future as well as present input prices rose by the same known percentage, the elasticity of demand for the commodity could be estimated from the decline in the inputs. Therefore the distinction between beneficial and harmful addiction is operational, these independently estimated commodity elasticities could be used, as in the text, to determine whether an addiction was harmful or beneficial.

in opposition to the private interests of the rulers for the time being. [p. 484]

The specific political behavior that contradicts "personal interest" theories is not clear from Mill's statement, nor is it much clearer in similar statements by others applied to firms or households. Obviously, stable behavior by (say) households faced with stable prices and incomes —or more generally a stable environment—is no contradiction since stability then is implied as much by personal interest theories as by custom and tradition. On the other hand, stable behavior in the face of changing prices and incomes might contradict the approach taken in this essay that assumes utility maximizing with stable tastes.

Nevertheless, we believe that our approach better explains when behavior is stable than do approaches based on custom and tradition, and can at the same time explain how and when behavior does change. Mill's "habits and modes of thought," or his "maxims and traditions which have descended," in our analysis result from investment of time and other resources in the accumulation of knowledge about the environment, and of skills with which to cope with it.

The making of decisions is costly, and not simply because it is an activity which some people find unpleasant. In order to make a decision one requires information, and the information must be analyzed. The costs of searching for information and of applying the information to a new situation are such that habit is often a more efficient way to deal with moderate or temporary changes in the environment than would be a full, apparently utility-maximizing decision. This is precisely the avoidance of what J. M. Clark termed the irrational passion for dispassionate rationality.

A simple example of economizing on information by the habitual purchase from one source will illustrate the logic. A consumer buys one unit of commodity X in each unit of time. He pays a price p_t at a time t. The choices he faces are:

1. To search at the time of an act of purchase to obtain the lowest possible price $\hat{p}_t$ consistent with the cost of search. Then $\hat{p}_t$ is a function of the amount of search s (assumed to be the same at each act of purchase):

$$\hat{p}_t = f(s),\ f'(s) < 0 \tag{10}$$

where the total cost of s is $C(s)$.

2. To search less frequently (but usually more intensively), relying between searches upon the outcome of the previous search in choosing a supplier. Then the price p_t will be higher (relative to the average market price), the longer the period since the previous search (at time t_0),

$$p_t = g(t - t_0),\ g' > 0$$

Ignoring interest, the latter method of purchase will have a total cost over period T determined by

1) K searches (all of equal intensity) at cost $K\,C(s)$.

2) Each search lasts for a period T/K, within which $r = T/K$ purchases are made, at cost $r\bar{p}$, where $\bar{p}$ is the average price. Assume that the results of search "depreciate" (prices appreciate) at rate δ. A consumer minimizes his combined cost of the commodity and search over the total time period; the minimizing condition is[11]

[11] The price of the ith purchase within one of the K search periods is $p_i = \hat{p}(1+\delta)^{i-1}$. Hence

$$\bar{p} = \frac{1}{r}\sum_{i=1}^{r} \hat{p}(1+\delta)^{i-1} = \hat{p}\,\frac{(1+\delta)^r - 1}{r\delta}$$

The total cost to be minimized is

$$TC = Kr\bar{p} + KC(s) = K\hat{p}\,\frac{(1+\delta)^r - 1}{\delta} + KC$$

By taking a second-order approximation to $(1+\delta)^r$, we get

$$TC = T\left\{\hat{p}\left[1 + \frac{(r-1)\delta}{2}\right] + \frac{C}{r}\right\}$$

Minimizing with respect to r gives

$$\frac{\partial TC}{\partial r} = 0 = T\left(\frac{\hat{p}\delta}{2} - \frac{C}{r^2}\right)$$

or

$$r = \sqrt{\frac{2C}{\delta\hat{p}}}$$

$$r = \sqrt{\frac{2C}{\delta \hat{p}}} \tag{11}$$

In this simple model with r purchases between successive searches, r is larger the larger the amount spent on search per dollar spent on the commodity $(C/\hat{p})$, and the lower the rate of appreciation of prices (δ). If there were full search on each individual act of purchase, the total cost could not be less than the cost when the optimal frequency of search was chosen, and might be much greater.

When a temporary change takes place in the environment, perhaps in prices or income, it generally would not pay to disinvest the capital embodied in knowledge or skills, or to accumulate different types of capital. As a result, behavior will be relatively stable in the face of temporary changes.

A related situation arises when an unexpected change in the environment does not induce a major response immediately because time is required to accumulate the appropriate knowledge and skills. Therefore, stable preferences combined with investment in "specific" knowledge and skills can explain the small or "inelastic" responses that figure so prominently in short-run demand and supply curves.

A permanent change in the environment, perhaps due to economic development, usually causes a greater change in the behavior of young than of old persons. The common interpretation is that young persons are more readily seduced away from their customs and traditions by the glitter of the new (Western?) environment. On our interpretation, young and old persons respond differently, even if they have the same preferences and motivation. To change their behavior drastically, older persons have to either disinvest their capital that was attuned to the old environment, or invest in capital attuned to the new environment. Their incentive to do so may be quite weak, however, because relatively few years remain for them to collect the returns on new investments, and much human capital can only be disinvested slowly.

Young persons, on the other hand, are not so encumbered by accumulations of capital attuned to the old environment. Consequently, they need not have different preferences or motivation or be intrinsically more flexible in order to be more affected by a change in the environment: they simply have greater incentive to invest in knowledge and skills attuned to the new environment.

Note that this analysis is similar to that used in the previous section to explain addictive behavior: utility maximization with stable preferences, conditioned by the accumulation of specific knowledge and skills. One does not need one kind of theory to explain addictive behavior and another kind to explain habitual or customary behavior. The same theory based on stable preferences can explain both types of behavior, and can accommodate both habitual behavior and the departures therefrom.

IV. Stability of Tastes and Advertising

Perhaps the most important class of cases in which "change of tastes" is invoked as an explanation for economic phenomena is that involving advertising. The advertiser "persuades" the consumer to prefer his product, and often a distinction is drawn between "persuasive" and "informative" advertising.[12] John Kenneth Galbraith is the most famous of the economists who argue that advertising molds consumer tastes:

> These [institutions of modern advertising and salesmanship] cannot be reconciled with the notion of independently determined desires for their central function is to create desires—to bring into being wants that previously did not exist. This is accomplished by the producer of the goods or at his behest.—Outlays for the manufacturing of a product are not more important in the strategy of modern business enterprise than outlays for the manufacturing of demand for the product. [pp. 155–56]

[12]The distinction, if in fact one exists, between persuasive and informative advertising must be one of purpose or effect, not of content. A simple, accurately stated fact ("I offer you this genuine $1 bill for 10 cents") can be highly persuasive, the most bizarre claim ("If Napoleon could have bought our machine gun, he would have defeated Wellington") contains some information (machine guns were not available in 1814)

We shall argue, in direct opposition to this view, that it is neither necessary nor useful to attribute to advertising the function of changing tastes.

A consumer may indirectly receive utility from a market good, yet the utility depends not only on the quantity of the good but also the consumer's knowledge of its true or alleged properties. If he does not know whether the berries are poisonous, they are not food; if he does not know that they contain vitamin C, they are not consumed to prevent scurvy. The quantity of information is a complex notion: its degree of accuracy, its multidimensional properties, its variable obsolescence with time are all qualities that make direct measurement of information extremely difficult.

How can this elusive variable be incorporated into the theory of demand while preserving the stability of tastes? Our approach is to continue to assume, as in the previous sections, that the ultimate objects of choice are commodities produced by each household with market goods, own time, *knowledge*, and perhaps other inputs. We now assume, in addition, that the knowledge, whether real or fancied, is produced by the advertising of producers and perhaps also the own search of households

Our approach can be presented through a detailed analysis of the simple case where the output x of a particular firm and its advertising A are the inputs into a commodity produced and consumed by households; for a given household:

$$(12) \qquad Z = f(x, A, E, y)$$

where $\partial Z/\partial x > 0$, $\partial Z/\partial A > 0$, E is the human capital of the houshold that affects these marginal products, and y are other variables, possibly including advertising by other firms. Still more simply,

$$(13) \qquad Z = g(A, E, y)x$$

where $\partial g/\partial A = g' > 0$ and $\partial^2 g/\partial A^2 < 0$. With A, E, and y held constant, the amount of the commodity produced and consumed by any household is assumed to be proportional to the amount of the firm's output used by that household.[13] If the advertising reaching any household were independent of its behavior, the shadow price of Z, the marginal cost of x, would simply be the expenditure on x required to change Z by one unit. From equation (13), that equals

$$(14) \qquad \pi_z = \frac{p_x}{g}$$

where p_x is the price of x.

An increase in advertising may lower the commodity price to the household (by raising g), and thereby increase its demand for the commodity and change its demand for the firm's output, because the household is made to believe—correctly or incorrectly—that it gets a greater output of the commodity from a given input of the advertised product Consequently, advertising affects consumption in this formulation not by changing tastes, but by changing prices. That is, a movement along a stable demand curve for commodities is seen as generating the apparently unstable demand curves of market goods and other inputs

More than a simple change in language is involved: our formulation has quite different implications from the conventional ones. To develop these implications, consider a firm that is determining its optimal advertising along with its optimal output. We assume initially that the commodity indirectly produced by this firm (equation (12)) is a perfect substitute to consumers for commodities indirectly produced by many other firms. Therefore, the firm is perfectly competitive in the commodity market, and could (indirectly) sell an unlimited amount of this commodity at a fixed commodity price. Observe that a firm can have many perfect substitutes in the commodity market even though few other firms produce the same physical product. For example, a firm may be the sole designer of jewelry that contributes to the social prestige of consumers, and yet compete fully with many other products that also contribute to prestige: large automobiles, expensive furs, fashionable clothing, elaborate parties, a respected occupation, etc.

If the level of advertising were fixed, there would be a one-to-one correspondence between the price of the commodity and the price of the firm's output (see equation (14)). If π_z were

[13]Stated differently, Z is homogeneous of the first degree in x alone

given by the competitive market, p_x would then also be given, and the firm would find its optimal output in the conventional way by equating marginal cost to the given product price. There is no longer such a one-to-one correspondence between π_z and p_x, however, when the level of advertising is also a variable, and even a firm faced with a fixed commodity price in a perfectly competitive commodity market could sell its product at different prices by varying the level of advertising. Since an increase in advertising would increase the commodity output that consumers receive from a given amount of this firm's product, the price of its product would then be increased relative to the fixed commodity price.

The optimal advertising, product price, and output of the firm can be found by maximizing its income

$$(15) \qquad I = p_x X - TC(X) - Ap_a$$

where X is the firm's total output, TC its costs of production other than advertising, and p_a the (constant) cost of a unit of advertising. By substituting from equation (14), I can be written as

$$(15') \qquad I = \pi_z^0 g(A)X - TC(X) - Ap_a$$

where π_z^0 is the given market commodity price, the advertising-effectiveness function (g) is assumed to be the same for all consumers,[14] and the variables E and y in g are suppressed. The first-order maximum conditions with respect to X and A are

$$(16) \qquad p_x = \pi_z^0 g = MC(X)$$

$$(17) \qquad \frac{\partial p_x}{\partial A} X = \pi_z^0 X g' = p_a$$

Equation (16) is the usual equality between price and marginal cost for a competitive firm, which continues to hold when advertising exists and is a decision variable. Not surprisingly, equation (17) says that marginal revenue and marginal cost of advertising are equal, where marginal revenue is determined by the level of output and the increase in product price "induced" by an increase in advertising. Although the commodity price is fixed, an increase in advertising increases the firm's product price by an amount that is proportional to the increased capacity (measured by g') of its product to contribute (at least in the minds of consumers) to commodity output.

In the conventional analysis, firms in perfectly competitive markets gain nothing from advertising and thus have no incentive to advertise because they are assumed to be unable to differentiate their products to consumers who have perfect knowledge. In our analysis, on the other hand, consumers have imperfect information, including misinformation, and a skilled advertiser might well be able to differentiate his product from other apparently similar products. Put differently, advertisers could increase the value of their output to consumers without increasing to the same extent the value of the output even of perfect competitors in the *commodity* market. To simplify, we assume that the value of competitors' output is unaffected, in the sense that the commodity price (more generally, the commodity demand curve) to any firm is not affected by its advertising. Note that when firms in perfectly competitive commodity markets differentiate their products by advertising, they still preserve the perfect competition in these markets. Note moreover, that if different firms were producing the same physical product in the same competitive commodity market, and had the same marginal cost and advertising-effectiveness functions, they would produce the same output, charge the same product price, and advertise at the same rate. If, however, either their marginal costs or advertising-effectiveness differed, they would charge different product prices, advertise at different rates, and yet still be perfect competitors (although not of one another)!

Not only can firms in perfectly competitive commodity markets—that is, firms faced with infinitely elastic commodity demand curves—have an incentive to advertise, but the incentive may actually be greater, the more competitive the commodity market is. Let us consider the

[14]Therefore, $p_x X = \pi_z^0 g \sum_{i=1}^{n} x_i$ where n is the number of households

case of a finite commodity demand elasticity.

The necessary conditions to maximize income given by equation (15′), if π_z varies as a function of Z, are

$$(18)\quad \frac{\partial I}{\partial X} = \pi_z g + X\frac{\partial \pi_z}{\partial Z}\frac{\partial Z}{\partial X}g - MC(X) = 0,$$

or since $Z = gX$, and $\partial Z/\partial X = g$,

$$(18')\quad \pi_z g\left(1 + \frac{1}{\epsilon_{\pi_z}}\right) = p_x\left(1 + \frac{1}{\epsilon_{\pi_z}}\right) = MC(X)$$

where ϵ_{π_z} is the elasticity of the firm's commodity demand curve. Also

$$(19)\quad \frac{\partial I}{\partial A} = X\frac{\partial p_x}{\partial A} - p_a = \pi_z\frac{\partial Z}{\partial A} + \frac{\partial \pi_z}{\partial Z}\cdot\frac{\partial Z}{\partial A}\cdot Z - p_a = 0$$

or

$$(19')\quad X\frac{\partial p_x}{\partial A} = \pi_z g' X\left(1 + \frac{1}{\epsilon_{\pi_z}}\right) = p_a$$

Equation (18′) is simply the usual maximizing condition for a monopolist that continues to hold when there is advertising.[15] Equation (19′) clearly shows that, given $\pi_z g'X$, the marginal revenue from additional advertising is greater, the greater is the elasticity of the commodity demand curve; therefore, the optimal level of advertising would be positively related to the commodity elasticity.

This important result can be made intuitive by considering Figure 1. The curve DD gives the firm's commodity demand curve, where π_z is measured along the vertical and commodity output Z along the horizontal axis. The firm's production of X is held fixed so that Z varies only because of variations in the level of advertising. At point e^0, the level of advertising is A_0, the product price is p_x^0, and commodity output and price are Z_0 and π_z^0, respectively. An increase in advertising to A_1 would increase Z to Z_1 (the increase in Z is determined by the given g' function). The decline in π_z induced by the increase in Z would be negatively related to the elasticity of the commodity demand curve: it would be less, for example, if the demand curve were $D'D'$ rather than DD. Since the increase in p_x is negatively related to the decline in π_z,[16] the increase in p_x, and thus the marginal revenue from the increase in A, is directly related to the elasticity of the commodity demand curve.[17]

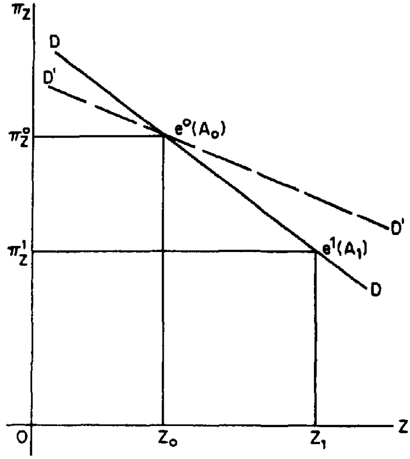


FIGURE 1

The same result is illustrated with a more con-

[15] If the level of advertising is held constant, Z is proportional to X, so

$$\epsilon_{\pi_z} = \frac{dZ}{Z}\Big/\frac{d\pi_z}{\pi_z} = \epsilon_{p_x} = \frac{dX}{X}\Big/\frac{dp_x}{p_x}$$

[16] Since $\pi_z g = p_x$,

$$\frac{\partial p_x}{\partial A} = \pi_z g' + g\frac{\partial \pi_z}{\partial A} > 0$$

The first term on the right is positive and the second term is negative. If g, g', and π_z are given, $\partial p_x/\partial A$ is linearly and negatively related to $\partial \pi_z/\partial A$

[17] Recall again our assumption, however, that even firms in perfectly competitive markets can fully differentiate their products. If the capacity of a firm to differentiate itself were inversely related to the elasticity of its commodity demand curve, that is, to the amount of competition in the commodity market, the increase in its product price generated by its advertising might not be directly related to the elasticity of its commodity demand curve.

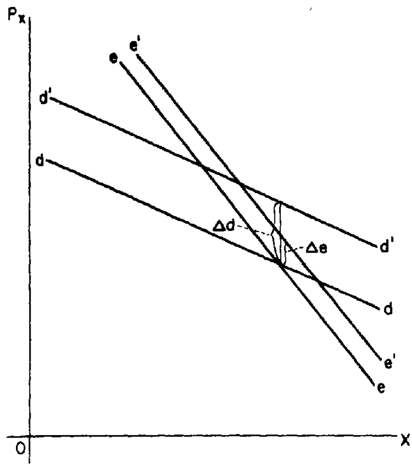


FIGURE 2

ventional diagram in Figure 2: the firm's product output and price are shown along the horizontal and vertical axes. The demand curve for its product with a given level of advertising is given by *dd*. We proved earlier (fn. 15) that with advertising constant, the elasticity of the product demand curve is the same as the elasticity of its commodity demand curve. An increase in advertising "shifts" the product demand curve upward to *d'd'*, and the marginal revenue from additional advertising is directly related to the size of the shift; that is, to the increase in product price for any given product output. Our basic result is that the shift is itself directly related to the elasticity of the demand curve. For example, with the same increase in advertising, the shift is larger from *dd* to *d'd'* than from *ee* to *e'e'* because *dd* is more elastic than *ee*.

This role of information in consumer demand is capable of extension in various directions. For example, the demand for knowledge is affected by the formal education of a person, so systematic variations of demand for advertisements with formal education can be explored. The stock of information possessed by the individual is a function of his age, period of residence in a community, and other variables, so systematic patterns of purchase of heavily and lightly advertised goods are implied by the theory.

V. Fashions and Fads

The existence of fashions and fads (short episodes or cycles in the consumption habits of people) seems an especially striking contradiction of our thesis of the stability of tastes. We find fashions in dress, food, automobiles, furniture, books, and even scientific doctrines.[18] Some are modest in amplitude, or few in their followers, but others are of violent amplitude: who now buys an ouija board, or a bustle? The rise and fall of fashions is often attributed to the fickleness of people's tastes. Herbert Blumer, the distinguished sociologist, gave a characteristic expression of this view:

> Tastes are themselves a product of experience, they usually develop from an initial state of vagueness to a state of refinement and stability, but once formed they may decay and disintegrate. . . .
>
> The fashion process involves both a formation and an expression of collective taste in the given area of fashion. The taste is intially a loose fusion of vague inclinations and dissatisfactions that are aroused by new experience in the field of fashion and in the larger surrounding world. In this initial state, collective taste is amorphous, inarticulate, and awaiting specific direction Through models and proposals, fashion innovators sketch possible lines along which the incipient taste may gain objective expression and take definite form. [p. 344]

The obvious method of reconciling fashion with our thesis is to resort again to the now familiar argument that people consume commodities, and only indirectly do they consume market goods, so fashions in market goods are compatible with stability in the utility function of commodities. The task here, as elsewhere, is to show that this formulation helps to illuminate our understanding of the phenomena under dis-

[18] "Fashion" indeed, does not necessarily refer only to the shorter term preferences Adam Smith says that the influence of fashion "over dress and furniture is not more absolute than over architecture, poetry, and music" (p 283)

cussion; we have some tentative comments in this direction.

The commodity apparently produced by fashion goods is social distinction: the demonstration of alert leadership, or at least not lethargy, in recognizing and adopting that which will in due time be widely approved. This commodity—it might be termed *style*—sounds somewhat circular, because new things appear to be chosen simply because they are new. Such circularity is no more peculiar than that which is literally displayed in a race—the runners obviously do not run around a track in order to reach a new destination. Moreover, it is a commendation of a style good that it be superior to previous goods, and style will not be sought intentionally through less functional goods. Indeed, if the stylish soon becomes inferior to the unstylish, it would lose its attractiveness.

Style, moreover, is not achieved simply by change: the newness must be of a special sort that requires a subtle prediction of what will be approved novelty, and a trained person can make better predictions than an untrained person. Style is social rivalry, and it is, like all rivalry, both an incentive to individuality and a source of conformity.

The areas in which the rivalry of fashion takes place are characterized by public exposure and reasonably short life. An unexposed good (automobile pistons) cannot be judged as to its fashionableness, and fashions in a good whose efficient life is long would be expensive. Hence fashion generally concentrates on the cheaper classes of garments and reading matter, and there is more fashion in furniture than in housing.

Fashion can be pursued with the purse or with the expenditure of time. A person may be well-read (i.e., have read the recent books generally believed to be important), but if his time is valuable in the market place, it is much more likely that his spouse will be the well-read member of the family. (So the ratio of the literacy of wife to that of husband is positively related to the husband's earning power, and inversely related to her earning power.)

The demand for fashion can be formalized by assuming that the distinction available to any person depends on his social environment, and his own efforts: he can be fashionable, give to approved charities, choose prestigious occupations, and do other things that affect his distinction. Following recent work on social interactions, we can write the social distinction of the ith person as

$$R_i = D_i + h_i \tag{20}$$

where D_i is the contribution to his distinction of his social environment, and h_i is his own contribution. Each person maximizes a utility function of R and other commodities subject to a budget constraint that depends on his own income and the exogenously given social environment.[19] A number of general results have been developed with this approach (see Becker), and a few are mentioned here to indicate that the demand for fashion (and other determinants of social distinction) can be systematically analyzed without assuming that tastes shift.

An increase in i's own income, prices held constant, would increase his demand for social distinction and other commodities. If his social environment were unchanged, the whole increase in his distinction would be produced by an increase in his own contributions to fashion and other distinction-producing goods. Therefore, even an average income elasticity of demand for distinction would imply a high income elasticity of demand for fashion (and these other distinction-producing) goods, which is consistent with the common judgement that fashion is a luxury good.[20]

If other persons increase their contributions to their own distinction, this may lower i's distinction by reducing his social environment. For distinction is scarce and is to a large extent simply redistributed among persons: an increase in one person's distinction generally requires a reduction in that of other persons. This is why people are often "forced" to conform to new fashions. When some gain distinction by paying

[19]The budget constraint for i can be written as

$$\Pi_{R_i} R + \Pi_z Z = I_i + \Pi_{R_i} D_i = S_i$$

where Z are other commodities, Π_{R_i} is his marginal cost of changing R, I_i is his own full income, and S_i is his "social income."

[20]Marshall believed that the desire for distinction was the most powerful of passions and a major source of the demand for luxury expenditures (see pp. 87–88, 106).

attention to (say) new fashions, they lower the social environment of others. The latter are induced to increase their own efforts to achieve distinction, including a demand for these new fashions, because an exogenous decline in their social environment induces them to increase their own contributions to their distinction.

Therefore, an increase in all incomes induces an even greater increase in i's contribution to his distinction than does an increase in his own income alone. For an increase in the income of others lowers i's social environment because they spend more on their own distinction; the reduction in his environment induces a further increase in i's contribution to his distinction. Consequently, we expect wealthy countries like the United States to pay more attention to fashion than poor countries like India, even if tastes were the same in wealthy and poor countries.

VI. Conclusion

We have surveyed four classes of phenomena widely believed to be inconsistent with the stability of tastes: addiction, habitual behavior, advertising, and fashions, and in each case offered an alternative explanation. That alternative explanation did not simply reconcile the phenomena in question with the stability of tastes, but also sought to show that the hypothesis of stable tastes yielded more useful predictions about observable behavior.

Of course, this short list of categories is far from comprehensive: for example, we have not entered into the literature of risk aversion and risk preference, one of the richest sources of *ad hoc* assumptions concerning tastes. Nor have we considered the extensive literature on time preference, which often alleges that people "systematically undervalue . . . future wants".[21] The taste for consumption in say 1984 is alleged to continue to shift upward as 1984 gets closer to the present. In spite of the importance frequently attached to time preference, we do not know of any significant behavior that has been illuminated by this assumption. Indeed, given additional space, we would argue that the assumption of time preference impedes the explanation of life cycle variations in the allocation of resources, the secular growth in real incomes, and other phenomena.

Moreover, we have not considered systematic differences in tastes by wealth or other classifications. We also claim, however, that no significant behavior has been illuminated by assumptions of differences in tastes. Instead, they, along with assumptions of unstable tastes, have been a convenient crutch to lean on when the analysis has bogged down. They give the appearance of considered judgement, yet really have only been *ad hoc* arguments that disguise analytical failures.

We have partly translated "unstable tastes" into variables in the household production functions for commodities. The great advantage, however, of relying only on changes in the arguments entering household production functions is that *all* changes in behavior are explained by changes in prices and incomes, precisely the variables that organize and give power to economic analysis. Addiction, advertising, etc. affect not tastes with the endless degrees of freedom they provide, but prices and incomes, and are subject therefore to the constraints imposed by the theorem on negatively inclined demand curves, and other results. Needless to say, we would welcome explanations of why some people become addicted to alcohol and others to Mozart, whether the explanation was a development of our approach or a contribution from some other behavioral discipline.

As we remarked at the outset, no conceivable expenditure of effort on our part could begin to exhaust the possible tests of the hypothesis of stable and uniform preferences. Our task has been oddly two-sided. Our hypothesis is trivial, for it merely asserts that we should apply standard economic logic as extensively as possible. But the self-same hypothesis is also a demanding challenge, for it urges us not to abandon opaque

[21]This quote is taken from the following longer passage in Bohm-Bawerk:

> We must now consider a *second* phenomenon of human experience—one that is heavily fraught with consequence. That is the fact that we feel less concerned about future sensations of joy and sorrow simply because they do lie in the future, and the lessening of our concern is in proportion to the remoteness of that future. Consequently we accord to goods which are intended to serve future ends a value which falls short of the true intensity of their future marginal utility. *We systematically undervalue our future wants and also the means which serve to satisfy them.* [p. 268]

and complicated problems with the easy suggestion that the further explanation will perhaps someday be produced by one of our sister behavioral sciences.

REFERENCES

G. S. Becker, "A Theory of Social Interaction," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1063–93.

H. C. Blumer, "Fashion," in Vol. V, *Int. Encyclo. Soc. Sci.*, New York 1968.

Eugen von Böhm-Bawerk, *Capital and Interest*, vol. 2, South Holland, IL 1959.

John K. Galbraith, *The Affluent Society*, Boston 1958.

M. Grossman, "The Economics of Joint Production in the Household," rep. 7145, Center Math. Stud. Bus. Econ., Univ. Chicago 1971.

Alfred Marshall, *Principles of Economics*, 8th ed., London 1923.

R. T. Michael and G. S. Becker, "On the New Theory of Consumer Behavior," *Swedish J. Econ.*, Dec. 1973, *75*, 378–96.

John S. Mill, *A System of Logic*, 8th ed., London 1972.

Adam Smith, *Theory of Moral Sentiments*, New Rochelle 1969.

Constant-Utility Index Numbers of Real Wages

By JOHN H. PENCAVEL*

Index numbers of workers' real wages are constructed by comparing observed changes in an index of consumer goods' prices with a measure of changes in their money wages. Four such index numbers, each based on slightly different definitions and drawn from different sources, are shown in Table 1. From 1939 to 1967 they record an increase in real wages of between 59 and 106.5 percent. As measures of the "true" standard of living of the typical worker, each is deficient in a number of respects. For instance, the consumer goods' price index used to deflate wages is the familiar base-weighted type which does not recognize that individuals will alter the composition of the basket of commodities they consume in response to relative price changes. It is well known that such a Laspeyres index overstates increases in the cost of maintaining a level of utility whenever consumers are induced by relative price movements to substitute among the commodities they purchase. Moreover, while the two Bureau of Labor Statistics (*BLS*) series on real spendable weekly earnings of production workers (as shown in columns (iii) and (iv) of Table 1) include an adjustment for federal income taxes and social security taxes, they do not discriminate between increases in weekly earnings that arise, on the one hand, through increases in hourly wage rates with hours worked constant and, on the other hand, through increases in hours worked with hourly wage rates fixed. An index of the ratio of average hourly earnings to consumer prices (as shown in column (i) of Table 1) is the familar technique for controlling for changes in hours worked. But presumably these movements in working hours themselves represent a response to changes in relative prices and in incomes, and a superior procedure is one that recognizes this choice in the allocation of time.

TABLE 1—INDEX NUMBERS OF REAL EARNINGS

Year	(i)[a]	(ii)[b]	(iii)[c]	(iv)[d]
1934	0.869	0.871	—	—
1939	1.000	1.000	1.000	1.000
1946	1.220	1 235	1 298	1 134
1950	1.327	1.401	1.364	1 219
1955	1 527	1.593	1.494	1.340
1960	1 687	1 744	1.555	1.412
1965	1 857	1.989	1 726	1 597
1967	1.919	2 065	1.717	1 593

Source: Bureau of Labor Statistics and Department of Commerce data

[a] An index of the ratio of average hourly earnings of production workers (as published by the *BLS*) to the consumer price index

[b] A Laspeyres real wage index defined as

$$(w_t/w_0)\left(\sum_{i=1}^{7} x_{0i}p_{ti} \Big/ \sum_{i=1}^{7} x_{0i}p_{0i}\right)^{-1}$$

and constructed from the Department of Commerce data used for the constant-utility index numbers formed below.

[c] The *BLS* index of real spendable (i e , after tax) weekly earnings for a production worker with three dependents.

[d] The *BLS* index of real spendable (i e , after tax) weekly earnings for a production worker with no dependents

This paper constructs index numbers of real income using estimated parameters that describe a particular form of the indirect utility function, namely, that derived from the linear expenditure system. These index numbers come from a model in which the length of work time, the size of labor income, and the pattern of consumer goods' purchases are decision variables. Nonlabor income enters as a determinant of working and consumption behavior. Constant-utility or true cost-of-living index numbers have been constructed before, but none has been derived from a system in which working decisions have been integrated with consumption decisions and

*Associate professor of economics, Stanford University. For their advice and assistance in the preparation of this paper, I should like to thank Orley Ashenfelter, Michael Metzger, Walter Oi, Nick Rau, Sherwin Rosen, David Weisberg, an anonymous referee, and the managing editor of this *Review*. This research received financial support from the Hoover Institution on War, Revolution, and Peace at Stanford University, the U.S. Department of Labor, and the Alfred P. Sloan Foundation

consequently none has distinguished the effects on utility of changes in wage rates from changes in nonlabor income.

The outline of the rest of the paper is as follows. Section I sketches the theoretical framework for the subsequent empirical analysis and raises the issue of what sort of constant-utility index numbers should be constructed. This question has not been confronted to date in the literature on cost-of-living index numbers. These numbers are usually constructed on the assumption that the individual's endowment of money income is invariant to changes in prices, whereas models of the allocation of time and labor supply are special cases of the general model of exchange in which the dollar amount available for expenditure on commodities is affected by changes in prices (in this case, changes in the wage rate). In this situation, either the wage rate or exogenous nonlabor income may form the basis for the construction of constant-utility index numbers. Solving for a constant-utility *price* (in this case, the wage rate) constitutes a departure from the universal practice in the literature on index numbers whose theorems are frequently conditional upon unitary *income* elasticities.[1] In Section II, the estimated parameters of an augmented Stone-Geary utility function are used to derive index numbers of wages which are contrasted with the information published by the Bureau of Labor Statistics. Some concluding observations for future research follow in Section III.

I. The Allocation of Time and Constant-Utility Index Numbers

One rationalization for the familar neoclassical commodity demand equations $x_j(p_1, \ldots, p_j, \ldots, p_n, I)$ involves characterizing an individual as making his consumption purchases such that his well-behaved utility function $U(x_1, \ldots, x_n)$ is maximized subject to a budget restraint $\Sigma_i p_i x_i = I$. Here, x denotes the commodities purchased, p the corresponding prices, and I stands for exogenous income. Upon substituting these commodity demand equations back into the utility function, the indirect utility function is obtained which relates the value of maximized utility to the arguments of the demand functions:[2] $V = V(p_1, \ldots, p_n, I)$. If observations on prices and income are available in each of two periods and if the form of the utility function is specified, then the indirect utility function may be evaluated for both periods and inferences drawn about changes in utility. Typically, to render the comparison meaningful in terms of dollars and cents, one determines the value of income in the second period which restores the first period's level of utility, after taking account of the consumer's response to the change in commodity prices. If this constant-utility income in year t is given by I_t^*, then I_t^*/I_0 is a true cost-of-living index and I_t/I_t^* is a true index of real incomes where the subscripts 0 and t designate the base year and the current year, respectively. It may be noted in passing that, instead of solving for the constant-utility level of income, in principle we could select one commodity price, say, p_k, and determine that price of x_k in the current period t which just compensates the individual for changes in all other prices and for changes in his income. In practice, we eschew this exercise and solve for the constant-utility level of income; there is a good intuitive understanding to determining how much income (rather than the price of, perhaps, brussel sprouts) has to be to restore an individual's base period utility.[3]

Suppose the individual's decision making is now expanded to include the division of his endowment of time (T) into working time (h) and nonmarket or leisure time (ℓ). Augmenting the utility function to admit nonmarket time as an argument and distinguishing labor and nonlabor income in the budget constraint, we proceed to determine the commodity demand and labor supply equations which upon substitution

[1]See the recent comprehensive survey by P. Samuelson and S. Swamy (1974).

[2]On the properties of the indirect utility function, see Lawrence Lau (1969).

[3]Nonetheless, there may be occasions on which a constant-utility price is sought. Thus the value of food stamps could be adjusted so as exactly to compensate consumers for changes in their money incomes and in all other prices. Of course, restricting the domain of relevant solutions to positive prices, there may be *no* price that exactly compensates the individual for changes in other prices and in income: a combination of rises in commodity prices and falls in income may mean that there is no positive p_k that restores base period utility.

into the utility function yield the indirect utility function:

$$V = V(p_1, \ldots, p_n, w, y)$$

where w stands for the wage rate and y for nonlabor income. This one-period model of decision making is itself a special case of an intertemporal allocation problem in which the optimal allocation of expenditures over time gives rise to saving or dissaving in any year. In this expanded scheme at least part of nonlabor income would be endogenous, the result of an optimal savings program, and perhaps only transfer payments from the government would be treated as given. The research reported in this paper, however, is restricted to the case in which all of nonlabor income is exogenous, and hence implicitly invokes all the assumptions (including certain, stationary price and wage expectations and intertemporal additivity) required to collapse the intertemporal problem into a sequence of one-period problems. In this context, the question arises of how best to measure utility changes between two years. A simple analogy with the consumer's allocation problem without the labor supply dimension might suggest determining that level of nonlabor income in year t (call it y_t^*) which reestablishes year 0's utility level after all commodity prices and the wage rate have changed. In effect, the Social Security Administration may be interpreted as making this sort of calculation when it determines the level of social security payments for working pensioners after the cost of living has changed. The zero homogeneity property of the indirect utility function means that a doubling of all prices and of the wage rate requires nonlabor income to be twice its base-year level if utility is to remain unchanged. But, instead of contrasting constant-utility nonlabor income with its base-year level, a comparison of y_t^* with actual nonlabor income in year t (y_t) provides an index of movements in real nonlabor income.[4]

Alternatively, suppose we solve for that wage rate in year t, w_t^*, which restores year 0's utility when all commodity prices and nonlabor income have changed. Again, invoking the homogeneity property of the maximized utility function, a doubling of all prices and of nonlabor income implies that the constant-utility wage rate w_t^* must be twice its base-year value. But, instead of comparing w_t^* with its base-year value, if we compare w_t^* with its actual value in year t, we arrive at a real wage index: when $w_t/w_t^* > 1$, wage rates are greater than necessary to maintain utility constant and hence the individual may be evaluated as being better off. This index w_t/w_t^* reflects the net effects of changes in all the arguments of the individual's indirect utility function and thus may be interpreted as a general index of the individual's welfare As an index of real wages, however, it has the unconventional feature of indicating, for instance, an increase in real wages when all commodity prices and wage rates are unchanged and yet nonlabor income has risen. Thus, a closer correspondence with traditional notions of real wage index numbers may be attained if a constant-utility wage rate is derived on the assumption that the individual worker's nonlabor income is fixed at its base-year level. Let w_t^{**} be that wage in year t which yields the same utility as in year 0 after commodity prices have changed but when nonlabor income is held fixed at its base-period level. Perhaps it resembles the wage rate that parties to collective bargaining contracts have in mind when arriving at a formula for linking movements in money wage rates to changes in commodity prices. Then w_t^{**}/w_0 is a "price-of-living" index while w_t/w_t^{**} is another real wage index number which this time includes the wage adjustments required to compensate the individual for the effects of changing commodity prices on fixed nonlabor income.[5]

II. Estimates of "True" Index Numbers of Real Wages and of Real Nonlabor Income

The empirical work on estimating systems of

[4]One possibility is to combine the wage rate and nonlabor income into a "full income" measure ($wT + y$) and determine constant-utility full income Though this would be an interesting exercise, it is not pursued here because of the essential arbitrariness in giving quantitative expression to the discretionary endowment of time (T)

[5]Of course, if in these experiments nonlabor income is permitted to rise but at less than the observed rate, the solved constant-utility wage rate will be bounded by w_t^* and w_t^{**}.

demand equations derivable from a given form of the utility function has tended to favor the use of the Stone-Geary (direct) utility function. This function permits the specification of unusually convenient estimating equations, and it appears to provide a good description of consumption patterns. Augmented to take account of the allocation of time between market and nonmarket activities, the Stone-Geary function has the following form:

$$U(x_1, \dots, x_n, l) = \sum_{i=1}^{n} B_i ln(x_i - \gamma_i) + \theta ln(l - \gamma_l)$$

where the parameters B_i, θ, γ_i, and γ_l obey the following restrictions: $\Sigma_i B_i + \theta = 1$; $(x_i - \gamma_i) > 0$ for $i = 1, \dots, n$; and $(l - \gamma_l) > 0$. If, in addition, each γ_i and γ_l is positive, they may be interpreted as "committed" or minimum subsistence quantities. The optimizing commodity demand and labor supply equations upon substitution into the utility function yield the following form for indirect utility:

(1) $$V(p_1, \dots, p_n, w, y) = \prod_{i=1}^{n} \left(\frac{B_i}{p_i}\right)^{B_i} \left(\frac{\theta}{w}\right)^{\theta} \left(y + w\gamma_h - \sum_{i=1}^{n} p_i \gamma_i\right)$$

where $\gamma_h = T - \gamma_l$ measures maximum feasible working hours.[6] The parameters of this function have recently been estimated by Michael Abbott and Orley Ashenfelter using aggregate time-series data for the *U.S.* economy for the years 1929 to 1967, and it is their estimates that are used in this paper for the construction of our constant-utility index numbers. The National Income and Product Accounts provided the source both for the consumption expenditure data (seven composite commodity groups were constructed) and for wage and salary income. To guarantee the exact satisfaction of the budget constraint, the conventional procedure is followed whereby nonlabor income is measured as the difference between total consumption expenditure and wage income.[7] Labor supply is measured by annual hours of work per employee which are taken from data compiled by Laurits Christensen and Dale Jorgenson. The ratio of total labor income to hours worked yields an hourly pretax wage rate and, upon multiplying this by the ratio of personal taxes to personal income, a series on the posttax hourly wage rate is derived.[8]

We omit a discussion here of the familiar problems encountered in deriving the parameter estimates of the utility function—the accuracy of the underlying observations, the requirements for the absence of aggregation bias, the conditions for the identification of these as commodity demand and labor supply functions, the estimation of a set of non-linear equations, and so forth—not because these are not critical issues, but because they have been discussed at length elsewhere and the contribution of this paper is to ascertain the implications of these parameters for the index numbers of earnings. The values of the parameters as estimated by Abbott and Ashenfelter with some implied elasticities are presented in Table 2. These elasticities are not strikingly at variance with the general drift of previous estimates that are derived from less restrictive functional forms: for instance, the common finding of durable goods being rela-

[6]Equation (1) may be viewed as a special case of that indirect utility function which arises in the general model of exchange in which an individual divides his endowment of given quantities of commodities into a part sold on the market and a part consumed himself. If in this Stone-Geary context $\bar{s}_i$ denotes the maximum feasible sales of commodity i, then the expression for the indirect utility function may be written

$$V = \prod_i \left(\frac{B_i}{p_i}\right)^{B_i} \left(y + \sum_i p_i \bar{s}_i\right)$$

In the special case in the text, fixed endowments are zero for commodities $i = 1, \dots, n$ so $\bar{s}_i = -\gamma_i$ and for the hours decision $\bar{s}_l = \gamma_l$.

[7]Thus any distinction between the effects of transfer payments on the one hand and of property income on the other hand on consumption and labor supply behavior is neglected here. Of course, some households will be in receipt of nonlabor income and yet be outside the labor force. This strains the tacit assumption that the wage-earning population is homogeneous with respect to the division of total income into its labor and nonlabor components.

[8]The imprecision in this adjustment for taxes arises both because average and marginal income tax rates differ, and because labor and nonlabor income tend to be taxed at different rates.

TABLE 2—POINT ESTIMATES OF PARAMETERS OF THE LINEAR EXPENDITURE SYSTEM AND MEASURES OF PRICE CHANGES IN THE DATA

Group	$\hat{\gamma}_i$ (and $\hat{\gamma}_h$)	$\hat{B}_i$ (and $-\hat{\theta}$)	Compensated elasticity	Uncompensated elasticity	Δp (and Δw)
Durables	−.922	.238	−1.290	−1.525	1.139
Food	699	.163	−.442	−.605	1.782
Clothing	.466	134	−.447	− 581	2.684
Other nondurables	.651	.0254	− 556	−.581	1 158
Housing services	.763	.0755	−.442	−.518	0.519
Transportation services	510	.0997	−.537	−.637	1.553
Other services	.424	.142	− 511	− 653	2.324
Allocation of hours	2357	− 121	.037	− 084	4 785

Notes. The $\hat{B}_i$ parameters are estimates of the marginal propensity to consume commodity i out of nonlabor income while $\hat{\theta}$ is the marginal propensity to "consume" leisure out of nonlabor income These are maximum likelihood estimates of the equations upon expressing the data in first-difference form and using numerical gradient methods to find the maximum of the likelihood function Working hours are measured as annual hours of work per employee Both the compensated own price (wage) elasticities and the uncompensated own price (wage) elasticities are evaluated at the sample means. Δp (and Δw) stands for the proportionate change in prices (and wage rates) from 1939 to 1967 as measured by these Department of Commerce data For further details on estimation method and data, consult Abbott and Ashenfelter.

tively own-price elastic and food relatively own-price inelastic in demand is replicated in this study. With respect to these estimates, perhaps we should mention that the constraint $(\gamma_h - h) > 0$ is violated for the (atypical ?) years 1929–33, 1937, and 1941–45. That is, the estimated value of γ_h falls short of actual hours worked. Therefore no index numbers are presented below for these years. The augmented linear expenditure system as estimated closely tracks the path of changes in consumption expenditure for the various commodity groups (including the "expenditure" on leisure) and compares favorably with the other systems of fitted commodity demand and labor supply equations.[9] With the parameters of this indirect utility function thus determined, the various index numbers of earnings may be computed.

Before proceeding with this, it is important to examine the underlying data in order to appreciate the sort of variation in the relative prices of the commodity groups (including leisure) over this period; if relative prices remained the same, there is little induced substitution among the commodities and thus little difference between a framework that permits intercommodity substitution and one that does not. The figures presented in the last column of Table 2 dismiss this possibility: the percentage increase in commodity prices between 1939 and 1967 ranges from a low of 52 for housing services to 268 for clothing, the latter exceeding the former by a factor of more than five. And once substitution between commodities and hours worked is permitted, the price variation of the arguments of the utility function ranges from a low of 52 percent for housing services to 479 percent for nonmarket time (or "leisure"). The substitution among commodities and leisure prompted by relative price and wage movements depends upon the form of the utility function, of course, but at a casual and impressionistic level the sort of relative price and wage variation revealed by these data would not typically be described as one of "stability."

First, consider solving for the wage rate in any year t (w_t^*) which restores the utility level of the base year after the prices of consumer goods and nonlabor income have changed at the observed rates. Given the form of the indirect utility function in equation (1), this constant-utility wage

[9]Thus, in this context, the addilog indirect utility function was judged as definitely inferior to the Stone-Geary (direct) utility function

TABLE 3—CONSTANT-UTILITY INDEX NUMBERS OF REAL WAGE RATES

Year	w_t^* (i)	$\sqrt{\text{var}(w_t^*)}$ (ii)	w_t^*/w_0 (iii)	w_t/w_t^* (iv)	w_t^{**} (v)	$\sqrt{\text{var}(w_t^{**})}$ (vi)	w_t^{**}/w_0 (vii)	w_t/w_t^{**} (viii)
1934	.370	.0010	.981	.867	359	.0013	952	.893
1939	.377	—	1.000	1.000	.377	—	1 000	1.000
1946	.542	.0142	1.438	1.309	.690	.0148	1.830	1.028
1950	.633	.0221	1.679	1 535	893	.0245	2.369	1.088
1955	785	0247	2.082	1.641	1.082	.0446	2.870	1.191
1960	.800	0376	2.122	2.010	1.262	0491	3.347	1.274
1965	.778	.0496	2.064	2.552	1.391	0544	3.690	1.427
1967	.878	0486	2 329	2 483	1.490	0661	3 952	1 463

Notes: The wage rates are measured in dollars per hour at work Columns (ii) and (vi) are estimates of the asymptotic standard errors of the constant-utility real wages These were calculated as follows Let r denote the column vector of first derivatives of, say, w_t^* with respect to the parameters γ_i, γ_h, B_i, and θ, i e ,

$$r' = \left[\frac{\partial w_t^*}{\partial \gamma_1}, \ldots, \frac{\partial w_t^*}{\partial \gamma_h}, \frac{\partial w_t^*}{\partial B_1}, \ldots, \frac{\partial w_t^*}{\partial \theta}\right]$$

If we evaluate these at the true parameter values, then the asymptotic variance of the estimator of w_t^* is var $(w_t^*) = r'\Omega r$ where Ω is the asymptotic variance-covariance matrix of the estimator of the parameters Substituting consistent estimators for the parameters yields a consistent estimator for var (w_t^*) as $\hat{r}'\hat{\Omega}\hat{r}$

rate is defined implicitly as follows:

$$\prod_i \left(\frac{p_{ti}}{p_{0i}}\right)^{B_i} \left(\frac{w_t^*}{w_0}\right)^{\theta} \left(\frac{y_0 + w_0\gamma_h - \sum_i p_{0i}\gamma_i}{y_t + w_t^*\gamma_h - \sum_i p_{ti}\gamma_i}\right) = 1$$

where the 0 subscript denotes the base year which, for purposes of comparison with the *BLS* series on real spendable earnings, we have designated as the year 1939.[10] The solved values for w_t^* with their accompanying asymptotic standard errors for selected years are listed in columns (i) and (ii) of Table 3.[11] The ratio of the constant-utility wage to the base year (1939) wage rate is given in column (iii) and this shows that in 1967 a wage rate some 2.3 times the wage in 1939 was necessary to maintain the base year's utility level. In fact, the actual wage rate was practically 2.5 times its constant-utility wage rate as shown in column (iv) of Table 3. This 150 percent increase in the real wage between 1939 and 1967 is, of course, more than double that recorded by the *BLS* real spendable weekly earnings series as shown in columns (iii) and (iv) of Table 1. It is also greater than that shown by a Laspeyres real wage index in column (ii) of Table 1 constructed from the same data that underline the estimates of w_t/w_t^* and defined as

$$(w_t/w_0)\left(\sum_i x_{0i}p_{ti} \Big/ \sum_i x_{0i}p_{0i}\right)^{-1}$$

This understatement of the *BLS* real spendable earnings series relative to our true real wage index w_t/w_t^* arises because the *BLS* neglects the utility increasing effects both of shorter working hours and of rising nonlabor income,[12] and because the *BLS* procedure of using a Laspeyres price index to deflate their earnings series tends to exaggerate the utility decreasing consequences of the rising prices of consumer goods. The consequences for our true real wage index of taking cognizance of rising nonlabor income are evident from column (v) of Table 3

[10] With nonzero committed quantities, the Stone-Geary indifference map is not homothetic to the origin and hence the constructed index numbers will not be independent of the choice of the base year.

[11] The complete series for these and other constant-utility measures are given in a table available upon request from the author.

[12] These data record a decline in annual hours of work per employee from 2334 hours in 1939 to 2126 in 1967 and a greater than threefold increase in nonlabor income over the same period.

which solves for the wage rate in year t (call it w_t^{**}) that restores 1939's utility when the prices of consumer goods have changed but when nonlabor income is maintained at its 1939 level. In the case of a Stone-Geary utility function, this constant-utility wage rate is defined implicitly in the following expression:

$$\prod_i \left(\frac{p_{ti}}{p_{0i}}\right)^{B_i} \left(\frac{w_t^{**}}{w_0}\right)^{\theta} \left(\frac{y_0 + w_0\gamma_h - \sum_i p_{0i}\gamma_i}{y_0 + w_t^{**}\gamma_h - \sum_i p_{ti}\gamma_i}\right) = 1$$

where, once again, the subscript 0 denotes the base year 1939. Holding nonlabor income constant in money terms implies, of course, a reduction in its value when the prices of consumer goods increase so that this constant-utility wage rate w_t^{**} so determined will necessarily be above w_t^* since 1939. That this is the case is evident from a comparison of columns (i) and (v). (The asymptotic standard errors of w_t^{**} are shown in column (vi) of Table 3.) The price of living index w_t^{**}/w_0 given in column (vii) of Table 3 shows that a wage rate some four times 1939's actual wage rate was required in 1967 as compensation for the price rises and thereby to restore 1939's utility level. In fact, the wage rate rose faster than would be required to maintain 1939's utility and, as indicated in column (viii), the actual wage rate in 1967 was about 1.5 times its constant-utility level w_t^{**}. Thus the increase from 1939 to 1967 in this true real wage index number, w_t/w_t^{**}, computed on the assumption of fixed nonlabor income, is much closer to that indicated by the *BLS* real spendable earnings series which omits consideration of nonlabor income altogether.

How much information concerning changes in these true real wage rates is conveyed by a given percentage change in the *BLS* real earnings series or in the ratio of average hourly earnings to the consumer price index? To answer this question, regression equations of the following form were specified:

$$\text{(2)} \qquad \Delta\left(\frac{w_t}{w_t^*}\right) = \alpha_{1i} + \beta_{1i}\Delta X_{i_t} + \epsilon_{1i_t} \qquad (i = 1,2,3)$$

$$\text{(3)} \qquad \Delta\left(\frac{w_t}{w_t^{**}}\right) = \alpha_{2i} + \beta_{2i}\Delta X_{i_t} + \epsilon_{2i_t} \qquad (i = 1,2,3)$$

where Δ is the proportionate change operator; X_1 is the *BLS* index of real spendable earnings for a worker with no dependents (as in column (iv) of Table 1); X_2 is the *BLS* index of real spendable earnings for a worker with three dependents (column (iii) of Table 1); and X_3 is the ratio of production worker average hourly earnings to the consumer price index (given in column (i) of Table 1). Clearly if $\alpha \approx 0$ and $\beta \approx 1$, then on average a given proportionate change in the readily available *BLS* measures of real earnings will be associated with the same proportionate change in the true real wage index.[13]

Before examining the regression results, observe that this is *not* a case in which a single government agency provides a set of data, and a relationship is being established between this series and another series derived from these *same* observations: the data on the right-hand side variables of equations (2) and (3) are compiled by the *BLS* while the Department of Commerce's National Income and Product Accounts supply the data underlying the true real wage index numbers. By way of analogy, an investigation of the association between the *BLS* Consumer Price Index *(CPI)* and the Department of Commerce's Implicit Price Deflator for Personal Consumption Expenditures *(PCE)*[14] found that quarterly changes in most *CPI* components do not forecast the same changes in the *PCE* components with a useful degree of precision. If these results were admitted as information

[13]To forestall confusion here, it is perhaps worth emphasizing that we are interested in the relationship between changes in the *measured* index numbers, that is, between changes in the X_i on the one hand, and changes in our constant-utility real wage index numbers on the other hand. The fact that both the right- and left-hand side variables almost certainly contain measurement error does not render the classical errors-in-variables model as the correct statistical characterization. For our task is *not* to ascertain the relationship between the error-free variables, but rather that between the measured index numbers. See Jack Triplett and Stephen Merchant.

[14]See Triplett and Merchant.

relevant to forming a prior judgement about the consequences of estimating equations (2) and (3) here, then presumably they would induce a scepticism that changes in the published *BLS* series of real earnings will provide a useful and accurate predictor of changes in these true real wage index numbers which are constructed from Department of Commerce data.[15]

The results from fitting equations (2) and (3) to data over the period 1947 to 1967 are shown in Table 4.[16] For the estimated equations (a), (b), and (c), turning to the t-distribution in the standard manner, the hypothesis $\alpha = 0$ is only rejected once at conventional levels of significance (namely for equation (3b)) and the hypothesis $\beta = 1$ is not rejected at all although the power of these tests in the case of the true index (w_t/w_t^*) is rather low. Moreover, the standard errors of estimate of equations (2a), (2b), and (2c) indicate that the published *BLS* earnings series is not very useful in forecasting changes in the true index (w_t/w_t^*): the standard errors of estimate which constitute a lower bound on the standard errors of forecast are of the order of 4.5 percentage points which is approximately the same as the observed standard deviation of $\Delta(w_t/w_t^*)$ over this postwar period. On the other hand, these *BLS* earnings data perform much better in predicting changes in the fixed nonlabor income true wage index (w_t/w_t^{**}): the standard errors of estimate of equations (3a) and (3b) are one-half the standard deviation of $\Delta(w_t/w_t^{**})$ and in a comparison involving the sums of squared residuals from equations (3c) and (3d) the hypothesis that changes in average hourly earnings and changes in the consumer price index are symmetrical (except for sign) in their effect on $\Delta(w_t/w_t^{**})$ cannot be rejected on a conventional F-test.[17] Hence little reliance can be placed in these *BLS* series on changes in real earnings *in any given year* for forecasting at all accurately movements in the true real wage index (w_t/w_t^*). On the other hand, they correspond much more closely to movements in the other true wage index (w_t/w_t^{**}) which is constructed on the assumption of a fixed nonlabor income.

Turn now to a constant-utility index number that involves determining in any year t the nonlabor income y_t^* necessary for a return to the utility surface in base year 0 after commodity prices and the wage rate have changed. In the Stone-Geary case, this constant-utility nonlabor income takes the following simple form:[18]

$$y_t^* = \left(\sum_i p_{ti}\gamma_i - w_t\gamma_h\right)$$

$$+ \left(y_0 + w_0\gamma_h - \sum_i p_{0i}\gamma_i\right)\prod_i \left(\frac{p_{ti}}{p_{0i}}\right)^{B_i}\left(\frac{w_t}{w_0}\right)^{\theta}$$

Not only may the first term in parentheses on the

[15]In fact, a Laspeyres real wage index formed from the Department of Commerce data is listed in column (ii) of Table 1. This suggests a growth of 107 percent in real wages over the 1939–67 period which is slightly greater than the 92 percent measured by the ratio of the *BLS* average hourly earnings to the consumer price index as given in column (i).

[16]The choice of estimating period was determined as follows. First, the published *BLS* series on real earnings commences in 1939. Second, the estimated value of the parameter γ_h falls short of actual hours worked for the war years 1941–45 and thereby does not satisfy one of the restrictions on the Stone-Geary utility function. For this reason, true real wage index numbers were not constructed for these years. Since we deal with between-year changes in these regressions, this restricts our observations to the years 1940 and 1947 to 1967. Whether 1940 is included or excluded makes no essential difference to the estimates in Table 4 so, to preserve a continuous and uninterrupted set of sample points, the years from 1947 to 1967 provided our observations for this exercise. Finally, for purposes of comparison with the equations using the *BLS* weekly earnings series as a right-hand variable, the equations with the ratio of average hourly earnings to the consumer price index are fitted to the same years (1947 to 1967) even though in this instance observations are available prior to 1939.

[17]To be explicit, equations (2d) and (3d) use the numerator (average hourly earnings) and the denominator (the consumer price index) of equations (2c) and (3c) as right-hand variables entered separately as determinants of changes in the true real wage index numbers as follows

$$\Delta\left(\frac{w_t}{w_t^*}\right) = \text{constant} + \delta_1\Delta\ (\text{average hourly earnings})_t$$

$$+ \delta_2\Delta\ (\text{consumer price index})_t + u_t$$

and similarly with $\Delta(w_t/w_t^{**})$ on the left-hand side. The F-test fails to reject the hypothesis that, with $\Delta(w_t/w_t^{**})$, $|\hat{\delta}_1| = |\hat{\delta}_2|$.

[18]If y_t^* were expressed as a fraction of base year's nonlabor income, an equation would be derived that closely resembles the familiar cost-of-living index from a Stone-Geary function when the utility implications of hours of

Table 4—Regressions of Changes in True Real Wages on Changes in Real Wages as Measured by the Bureau of Labor Statistics, 1947–1967

Equation number	Right-hand variable(s)	Left-hand variable	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\delta}_2$	R^2	SEE	DW
(2a)	1. Proportionate change in the *BLS* index of real spendable earnings for a worker with no dependents	$\Delta(w_t/w_t^*)$	.013 (.012)	1.167 (.457)			.256	.041	1.84
(3a)		$\Delta(w_t/w_t^{**})$	.0034 (.0027)	.832 (.105)			.768	.0094	1.97
(2b)	1. Proportionate change in the *BLS* index of real spendable earnings for a worker with 3 dependents	$\Delta(w_t/w_t^*)$	.021 (.012)	.789 (.515)			.110	.045	1.63
(3b)		$\Delta(w_t/w_t^{**})$	.0054 (0025)	.858 (.107)			.771	.0093	2.26
(2c)	1 Proportionate change in the ratio of average hourly earnings to the consumer price index	$\Delta(w_t/w_t^*)$	.014 (018)	828 (.677)			.073	.046	1.56
(3c)		$\Delta(w_t/w_t^{**})$	−.0051 (0045)	1.014 (171)			.648	.012	2.21
(2d)	1. Proportionate change in production worker average hourly earnings, and	$\Delta(w_t/w_t^*)$	.057 (.032)		−.334 (963)	−.329 (.714)	.183	.044	1.45
(3d)	2 Proportionate change in consumer price index	$\Delta(w_t/w_t^{**})$	0037 (.0081)		−.777 (.245)	−.919 (.182)	.688	.011	2.40

Note: Estimated standard errors are shown in parentheses beneath the estimated regression coefficients *SEE* is the standard error of estimate for the regression equation and *DW* is the Durbin-Watson statistic For these years (1947–67), $\Delta(w_t/w_t^*)$ possesses a mean of .032 and a standard deviation of 046; the corresponding figures for $\Delta(w_t/w_t^{**})$ are .017 and .019, respectively

Table 5—Constant-Utility Nonlabor Income

Year	y_t^* (i)	$\sqrt{\text{var}(y_t^*)}$ (ii)	y_t^*/y_t (iii)	$1+\left[\frac{(y_t - y_t^*)}{y_t}\right]$ (iv)	y_t/y_t^* (v)
1934	632 6	5 30	1 190	.810	.840
1939	553.9	—	1 000	1 000	1 000
1946	513.2	23.91	595	1 404	1.681
1950	386 1	33 53	352	1 648	2.841
1955	110 0	52 64	093	1.906	10 440
1960	−194.6	75.81	−.128	2.128	−7.813
1965	−740.0	104.31	− 403	2 403	−2.481
1967	−954.6	120.50	−.518	2.518	−1 931

Note: *Nonlabor income is measured in annual dollars per employee. The estimated asymptotic* standard errors of y_t^* are listed in column (ii) and are computed by the method outlined in the notes of Table 3

work are ignored. This equation would be a weighted average of an arithmetic and a geometric index where the weights depend upon the ratio of "committed" *net* purchases to nonlabor income.

right-hand side be negative (i.e., $\Sigma p_{ti}\gamma_i < w_t\gamma_h$), but in addition it may be larger in absolute magnitude than the rest of the expression. This negative value for y_t^* would mean the

utility-increasing consequences of wage rates rising relative to commodity prices had been such that our representative individual is required to make net payments rather than enjoy net receipts of nonlabor income to restore the base year's utility. The series on y_t^* in column (i) of Table 5 indicates that from the late 1950's this is exactly what has happened. By 1967 the adjustment in nonlabor income needed to restore 1939's utility level entails some \$950 being taxed away from our representative individual. In fact, these data record that nonlabor income was \$1841 per employee in 1967 so that the *relative* difference between actual and constant-utility nonlabor income over the 1939–67 period increased on the order of 152 percent as given in column (iv).[19]

III. Further Research

Since these are the first constant-utility real wage index numbers known to the author, there is no opportunity to contrast these with others derived from a different specification of the utility function. This would represent an obvious and natural extension of the work described in this paper. A comparison of these Stone-Geary based index numbers with those from some other representation of preferences would provide some idea of the robustness of the true indexes constructed here. Also, there is another class of constant-utility index numbers that apply when one or more goods are effectively rationed. For instance, if working hours are rationed by employers, the evaluation of the indirect utility function is conditional upon the number of constrained hours of work.[20] The construction of this class of index numbers involves reestimating the whole system of commodity demand and labor supply equations on the assumption that working hours are preallocated by employers. Finally, no distinction has been made in this paper between the individual worker and the household. Yet, if it is the latter that forms both the relevant decision-making unit and the unit of concern for social policy, then the labor market opportunities of each member of the household must be incorporated into a correct evaluation of changes in real income. The results of some research on this issue will be reported at a later date.

[19]The definition of this index is obviously different from the preceding ones involving the wage rate. The reason for this practice here is to avoid the difficulties of interpretation when y_t is expressed as a fraction of negative values of y_t^* (see col. (v) of Table 5).

[20]Thus, in this Stone-Geary case, the ratio of the actual wage to the wage (call it w_t^*) that restores base year's utility when commodity prices and nonlabor income have changed and when hours are preallocated to be $\bar{h}$ is

$$\frac{w_t}{w_t^*} = \left(\frac{w_t\bar{h}_t}{\sum_i p_{ti}\gamma_i - y_t}\right) + \left(\frac{w_t\bar{h}_t}{y_0 + w_0\bar{h}_0 - \sum_i p_{0i}\gamma_i}\right)\prod_i \left(\frac{p_{0i}}{p_{ti}}\right)^{(\theta_i/(1-\theta))}\left(\frac{\gamma_h - \bar{h}_t}{\gamma_h - \bar{h}_0}\right)^{(\theta/(1-\theta))}$$

REFERENCES

M. Abbott and O. Ashenfelter, "Labor Supply, Commodity Demand, and the Allocation of Time," *Rev. Econ. Stud.*, Oct. 1976, *43*.

L. R. Christensen and D. W. Jorgenson, "U.S. Real Product and Real Factor Input, 1929–1967," *Rev. Income and Wealth*, Mar. 1970, *16*, 19–50.

L. J. Lau, "Duality and the Structure of Utility Functions," *J. Econ. Theory*, Dec. 1969, *1*, 374–96.

P. A. Samuelson and S. Swamy, "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis," *Amer. Econ. Rev.*, Sept. 1974, *64*, 566–93.

J. E. Triplett and S. M. Merchant, "The CPI and the PCE Deflator: An Econometric Analysis of Two Price Measures," *Ann. Econ. Soc. Meas.*, Summer 1973, *2*, 263–82.

J. R. N. Stone, "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," *Econ. J.*, Sept. 1954, *64*, 511–27.

U.S. Department of Commerce, *The National Income and Product Accounts of the United States 1929–1965*, Washington 1966.

———, *U.S. National Income and Product Accounts, 1964–76*, Washington 1971.

U.S. Bureau of Labor Statistics, *Employment and Earnings*, various issues.

Unanticipated Money Growth and Unemployment in the United States

By ROBERT J. BARRO*

The hypothesis that forms the basis of this empirical study is that only unanticipated movements in money affect real economic variables like the unemployment rate or the level of output. This hypothesis is explicit in "rational expectation" monetary models, such as those of Robert Lucas (1972, 1973), Thomas Sargent and Neil Wallace, and the author (1976a). However, the proposition that only the unanticipated part of money movements has real effects is clearly more general than the specific setting of these models.

In order to implement and test the hypothesis empirically, it is necessary to quantify the notions of anticipated and unanticipated money movements. Accordingly, the first part of the analysis specifies a simple model of the money growth process. The variables that turn out empirically to have a systematic effect on *U.S.* money growth, using annual observations from 1941 to 1973, are a measure of federal government expenditure relative to "normal," a lagged unemployment rate, and two lagged values of money growth. Anticipated money growth is then viewed as the prediction that could have been obtained by exploiting the systematic relation between money growth and this set of independent variables.

The measure of unanticipated money growth—actual growth less the anticipated portion—that is obtained in Section I is used in Section II as an explanatory variable for the unemployment rate—the real economic variable that is focused on in the present study. Over the 1946–73 period, the contemporaneous and two annual lag values of unanticipated money growth turn out to have effects that are significantly negative on unemployment. Further, the hypothesis that only the unanticipated part of money expansion influences unemployment receives strong support from some empirical tests.

The final sections discuss unemployment predictions, implications for policy, and some possibilities for extension of the research.

I. Analysis of Money Growth

A. *Setup of the Equation*

The money growth rate equation used in this study applies to annual observations for the 1941–73 period. The equation includes the following variables: a measure of federal government expenditure relative to normal, the lagged unemployment rate, and two lagged values of money growth. The government expenditure variable captures an aspect of the revenue motive for money creation. In my 1976b paper I describe a theoretical model in which an exogenous level of government expenditure is financed by a combination of taxes and money issue (Extensions to include public debt and nongovernment money do not alter the main conclusions.) Each method of finance involves administrative and other deadweight costs that increase, *ceteris paribus*, at an increasing rate with the amount of revenue raised by that method. However, the costs of raising a given amount of revenue by either method are assumed to decline with an increase in national income. In addition, the costs associated with taxation are assumed to depend negatively on the amount of fixed "capital" that has been accumulated in tax-raising capacity. For example, the setting up of an income tax and of an institutional apparatus for administering this tax are viewed as increases in tax-raising capital that reduce the collection costs imputed by the government to any particular amount of

*University of Rochester. The National Science Foundation has supported this research. I have benefited from comments on earlier drafts by Jack Carr, Bob Hodrick, Pieter Korteweg, Bob Lucas, Michael Parkin, and Chris Sims.

TABLE 1—VALUES OF MONEY GROWTH AND UNEMPLOYMENT

	DM	*D̂M*	*DMR*	*U*	*Û*	*U-Û*	*UNAT*
1939	.114	—	—	—	—	—	—
1940	.151	—	—	.095	—	—	—
1	.160	.166	−.007	.058	—	—	—
2	180	212	−.032	029	—	—	—
3	.265	201	.064	.015	—	—	—
4	162	192	−.031	.010	—	—	—
1945	.150	158	−.008	.016	—	—	—
6	.068	055	.013	.037	.039	−.002	.034
7	.047	.038	.009	.038	043	−.005	.051
8	004	.017	−.012	.037	.044	−.007	.048
9	−.010	.013	−.023	.057	.053	.004	.042
1950	.026	.006	019	.052	.059	−.007	.048
1	.044	.026	.018	.031	.031	.000	.039
2	.049	.037	.012	.028	025	.003	.036
3	.024	.041	−.017	.027	.032	−.005	.035
4	.015	.024	− 008	.052	.044	.008	.036
1955	.031	027	004	.042	.043	−.001	.037
6	.012	021	− 009	.039	042	−.003	.040
7	.005	023	−.018	.041	049	−.008	.041
8	.012	.020	−.007	.065	.054	.011	.041
9	.037	.030	.007	.053	046	.007	.041
1960	−.001	.030	−.031	053	046	007	.041
1	021	.033	− 013	.065	.062	.003	.042
2	.022	.033	−.012	.053	059	−.006	.042
3	.029	.033	−.004	.055	053	002	043
4	.039	.035	.004	050	047	.003	.044
1965	042	.037	.005	.043	041	.002	.043
6	044	042	002	.037	.038	−.001	.042
7	.039	.043	−.004	.036	042	− 006	.043
8	.068	042	.026	.034	040	− 006	.044
9	.061	.041	020	.034	.030	.004	.044
1970	044	047	− 004	.047	.046	.001	.064
1	.067	049	018	.057	.054	003	.062
2	.063	.056	006	.054	.048	.006	060
3	.071	.061	.010	048	.049	− 001	.059
4	.055	.056	−.001	.055	.056	− 001	.064
1975	.042	062	−.020	.083	071	012	065
6	—	065	—	—	081	—	.065
7	—	.065	—	—	068	—	063
8	—	.069	—	—	.061	—	061

Notes: $DM_t = log(M_t) - log(M_{t-1})$, where M is an annual average of M_1 from the *Federal Reserve Bulletin*, $\hat{DM}$ is the estimated value from equation (2), $DMR = DM - \hat{DM}$; U is the annual average unemployment rate (data are given in the *Economic Report of the President*), based on the total labor force, which includes military personnel. Data for 1940–43 were adjusted for treatment of government "emergency workers," as discussed in Michael Darby; $\hat{U}$ is an estimated value from equation (4); *UNAT* is derived from equation (4) with all *DMR* values set equal to zero

Values of $\hat{DM}$ for 1976–78 are based on the value $FEDV = .18$. The 1977–78 values of *DM* use the value of $\hat{U}$ from the preceding year. The values of $\hat{U}$ (and *UNAT*) subsequent to 1975 are based on $DMR = 0$ for 1976 and beyond, $MIL = 0$, and the *MINW* values indicated in Table 2.

revenue raised by taxes. The amount of capital invested in the tax-raising "industry" is, itself, endogenous to the model, but adjustment-type costs associated with changes in taxing capacity imply that the current amount of capital will depend on "long-run" values of such variables as government expenditure and national income, rather than simply on the current values.

The breakdown of the total government budget between the two revenue components—and therefore the rate of money growth—is determined to minimize the total costs associated with raising revenue.[1] With a fixed amount of tax-

[1]Edmund Phelps has a theoretical discussion of inflation within a public finance context.

raising capital—which depends, among other things, on the long-run level of government expenditure—an increase in the current government budget leads to increases in both types of revenue and, hence, to an increase in the growth rate of money. This type of response would apply especially to periods of wartime during which there are sharp, temporary increases in the size of the government budget. The long-run response of money to an increase in the government budget—for example, to the secular growth of federal government spending relative to *GNP* that occurred from the 1930's to the 1960's—would be different, because it would lead also to an increase in tax-raising capital. Since "permanent" expansions in the share of *GNP* that is absorbed by government lead to an expansion of taxing capacity, it is possible—depending on the form of the "production function" that generates tax revenues—that no change in the money growth rate would occur. In this situation it is only increases in government expenditure relative to normal that would induce monetary expansion.

Specifically, my equation for money growth uses the variable

$$FEDV \equiv log(FED) - [log(FED)]^*$$

where *FED* is real expenditure of the federal government and $[log(FED)]^*$ refers to the normal value of this expenditure. Empirically, $[log(FED)]^*$ was generated from the adaptive formula[2]

$$[log(FED)]_t^* = \beta[log(FED)]_t + (1-\beta)[log(FED)]_{t-1}^*$$

that is, $[log(FED)]^*$ is an exponentially declining distributed lag of $log(FED)$. The equation reported below uses the value of the adaptation coefficient, $\beta = 0.2$ per year (see fn. 5). Values of *FEDV* corresponding to this value of β are tabulated from 1941 to 1975 in Table 2.

[2]It would be preferable to generate $[log(FED)]^*$ from a prediction relation based on the time-series properties of $log(FED)$. Costs of adjustment in taxing capacity would then also have to be taken into account in specifying the reaction of money growth to optimal predictions of future values of $log(FED)$. I have not yet proceeded along these lines.

In the empirical analysis I test the hypothesis that it is only the difference between $log(FED)$ and $[log(FED)]^*$ that influences money expansion, with no separate effect of the level of fed-

TABLE 2—VALUES OF THE FEDERAL EXPENDITURE, MILITARY PERSONNEL, AND MINIMUM WAGE VARIABLES

	$FEDV(\beta = 2)$	*MIL*	*MINW*
1941	.803	—	—
2	1.356	—	—
3	1 369	—	—
4	1.161	—	—
1945	812	—	—
6	−.131	105	.228
7	− 338	012 (048)	203
8	− 196	.022 (.044)	.191
9	− 016	048	.180
1950	− 033	.049	.323
1	199	092	301
2	307	106	.284
3	.303	105	269
4	151	099	.258
1955	.090	.090	.248
6	.089	083	307
7	123	081	298
8	167	075	283
9	139	073	.273
1960	.116	.071	.262
1	157	071	283
2	179	.077	328
3	157	073	325
4	143	072	334
1965	136	071	325
6	204	.079	.315
7	245	086	392
8	248	087	.426
9	195	085	.421
1970	.173	0 (075)	402
1	165	0 (065)	.367
2	188	0 (056)	344
3	172	0 (052)	322
4	154	0 (048)	408
1975	.195	0 (.046)	426
6	(18)	(0)	(42)
7	(18)	(0)	(39)
8	(18)	(0)	(.36)

Notes $FEDV(\beta = .2)$, *MIL*, and *MINW* are the federal expenditure, military personnel, and minimum wage variables, as defined in the text The military values shown in parentheses for certain years are the actual ratios of military personnel to the male population aged 15–44, ignoring the absence of a selective draft for all or part of those years.

The FEDV value of 18, shown in parentheses for 1976–78, corresponds to the average value over 1960–75. The *MINW* value shown in parentheses for 1976 is an estimated value that takes account of the rise in the nominal minimum wage on January 1, 1976. The 1978 value of .36 is the sample average over 1960–75 For purposes of predicting unemployment for 1977 and beyond, it is assumed that the *MINW* variable will fall over a two-year period from its 1976 value to the average value of .36.

eral expenditure. This hypothesis is supported by the empirical evidence over 1941 to 1973, so that the main analysis includes only the *FEDV* variable.

The money growth equation also includes a measure of lagged unemployment. A positive response of money growth to this variable could reflect two elements. First, there could be a countercyclical policy response of money to the level of economic activity. (The subsequent analysis has important implications for the efficacy of this type of policy.) Second, a decline in real income lowers holdings of real balances, which would reduce the amount of government revenue from money issue for a given value of the money growth rate. As shown by the author (1976b), the optimal response to a decline in income below its normal level would be an increase in the money growth rate. My empirical analysis does not separate out these two possible sources of countercyclical money response.

Finally, the money growth equation includes two lagged values of money growth as "explanatory" variables. Presumably, these lagged dependent variables pick up any elements of serial dependence or lagged adjustment that have not been captured by the other independent variables.

The form of the systematic part of the money growth equation is

$$(1)\quad DM_t = \alpha_0 + \alpha_1 DM_{t-1} + \alpha_2 DM_{t-2} + \alpha_3 FEDV_t + \alpha_4 UN_{t-1}$$

where M_t is an annual average of M_1 (see fn. 18 below on the money definition), $DM_t \equiv log(M)_t - log(M)_{t-1}$ measures the annual average money growth rate, $FEDV_t \equiv log(FED)_t - [log(FED)]_t^*$, as defined above,[3] and $UN_{t-1} \equiv log(U/(1-U))_{t-1}$, where U is the annual average unemployment rate in the total labor force (which includes military personnel). The form in which the unemployment rate enters corresponds to the form of the unemployment equation given below.

[3]Data on federal government expenditures are from the *Economic Report of the President*, various issues. The figures on the nominal federal budget were divided by the *GNP* deflator (1958 = 1.0).

B. *Estimated Equation*

The estimated money growth equation for the 1941–73 period is, with standard errors in parentheses,[4]

$$(2)\quad DM_t = \underset{(.031)}{.087} + \underset{(.15)}{0.24}DM_{t-1} + \underset{(.13)}{0.35}DM_{t-2} + \underset{(.015)}{.082}\cdot FEDV_t + \underset{(.010)}{.027}\cdot UN_{t-1},$$

$$R^2 = .90,\ \hat{\sigma} = .020,\ D.W. = 2.39,$$
(sample average of $DM = .057$)

where $\hat{\sigma}$ is the standard error of estimate.

Consider, first, the coefficient on the federal expenditure variable, *FEDV*.[5] The estimated value of .08 implies that a 10 percent increase in real federal expenditure—holding fixed the normal expenditure and lagged values of *DM* —would raise DM_t by $\frac{8}{10}$ of a percentage point per year. Historically, the extreme values of *FEDV* have occurred during and just after wartime periods. For example, the 1943 value of $FEDV = 1.37$ implies (with DM_{t-1} and DM_{t-2} held fixed, so that .08 is the applicable coefficient) that DM_t would be 11 percentage points per year higher than when FEDV is zero, while the 1947 value of $FEDV = -.34$ implies that DM_t would be 3 percentage points per year less than otherwise.[6]

[4]A measure of the contemporaneous or lagged value of the federal government deficit relative to *GNP* is insignificant when added to equation (2). A lagged value of the inflation rate (based on the *GNP* deflator) or of the interest rate on prime commercial paper is also insignificant.

[5]Based on the fit of the money growth equation, the maximum likelihood estimate of the adaptation coefficient β is in the interval between 0.15 and 0.20, with an asymptotic 95 percent confidence interval of (0.1, 0.4). Since the unemployment results showed little sensitivity to variations in β over the interval from 0.15 to 0.30, I have limited the reported results to the case of $\beta = 0.20$.

[6]Note, however, that *FEDV* has not been normalized to make the long-run average value equal to zero. Since the normal value of government expenditure is generated by a distributed lag of actual values, secular growth of the public sector implies that the typical measured value of *FEDV* will be positive. It turns out that constant growth of real expenditures at rate g would generate a *FEDV* value of $g(1-\beta)/\beta$, which equals $4g$ at $\beta = 0.2$. From 1949 to 1973 the average annual growth rate g is .050, so that the corresponding "long-run average" value of *FEDV* is .20. However, growth of the public sector at 5 percent per year would not seem to be permanently sustainable.

The hypothesis that government expenditure enters into the determination of money growth only as the difference $FEDV \equiv log(FED) - [log(FED)]^*$ has been tested by entering $FEDV$ and $log(FED)$ separately into the DM equation. The estimated coefficient of $log(FED)$ (.010, standard error = .010) differs insignificantly from zero, and there is little change in the estimated coefficients on the other variables. (The results are similar if FED is measured as a ratio to a trend value of real GNP.) Accordingly, the results are consistent with the view that only temporary movements in federal expenditure stimulate monetary expansion.

Consider next the coefficient on the lagged unemployment variable. The estimated value of .03 implies that a 10 percent increase in U—that is, an increase by $\frac{1}{2}$ percentage point starting from $U = 5$ percent—would imply a reaction of next year's money growth by about $\frac{3}{10}$ of a percentage point per year. Hence, an increase by 1 percentage point in the unemployment rate induces an increase in next year's money growth rate by about $\frac{6}{10}$ of a percentage point per year.[7]

Finally, the regression results indicate persistence effects with an estimated DM_{t-1} coefficient of 0.24 and an estimated DM_{t-2} coefficient of 0.35. If the error term in equation (2) is assumed to follow a first-order Markov process $u_t = \rho u_{t-1} + \epsilon_t$, the maximum likelihood estimate of ρ is $-.35$ (the estimated coefficient on DM_{t-1} is then 0.45 and that on DM_{t-2} is 0.21). However, the estimated value of ρ differs insignificantly from zero at the 5 percent level—the asymptotic chi-square value is 2.6 with a critical value of 3.8. Since the inclusion of a nonzero value for ρ also has a negligible impact on the subsequent analysis of unemployment, I have limited the main analysis to the case where $\rho = 0$.

[7]From either the countercyclical policy or optimal revenue-raising viewpoints, it would seem preferable to enter the unemployment variable relative to its perceived long-run average value. The results on unemployment over 1946 to 1973, below, suggest that the principal movement in this long-run average value may have occurred since 1970. However, I have not yet attempted to adjust the unemployment variable along these lines.

A notable aspect of the estimated DM equation is that it implies a normal, or long-run average, money growth rate. For a given value of the constant term and the federal expenditure and unemployment variables, the equation specifies the mean value of DM (both in a short-run sense conditioned on given values of DM_{t-1} and DM_{t-2}, and also in a long-run unconditional sense). For example, if the unemployment rate is 4.3 percent, the average estimated "natural rate" during the 1960's (as discussed below), and if the $FEDV$ variable takes on an "average" value of .20 (see fn. 6), the implied long-run mean value of DM is 4.4 percent per year.

If the DM equation had contained a distributed lag of past DM values with the lag coefficients summing to one, as is true in the adaptive expectations formula developed by Phillip Cagan, then the model would not have the property of possessing a natural or long-run mean value of the money growth rate. Presumably, the formulation with lag weights summing to unity would provide a satisfactory framework for predicting DM only if money growth were, in fact, generated by a nonstationary process (for example, a random walk, or a random walk observed with error as in John Muth) for which a long-run mean did not exist. If the money growth process is stationary, it would not be expected that the sum of the lag weights in money growth predictions would equal one. It follows that any implicit tests of expectation formation concerning money growth that are based on lag weights summing to one would not generally be meaningful—a point that was made in a general context by Sargent (1971).

C. *Prior Predictions of Money Growth*

For the present analysis, the purpose of fitting a money growth equation is to obtain a division of money growth into anticipated and unanticipated components. The theoretical proposition is then that only the unanticipated part of money growth will influence unemployment. There is a basic problem to consider in using an estimated money growth equation to specify the concept of anticipated money growth. Consider the formulation of this anticipation for date t, $\widehat{DM}_t$. This

anticipation could be based on information that was available up to date t − 1, and might also include partial information applicable to date t. However, $\widehat{DM}_t$ should not be based on any information that becomes available only after date t. For example, if the estimated values from the *DM* regression for the 1941–73 period were used to obtain $\widehat{DM}$ for 1950, then information subsequent to 1950 would be used to "predict" that year's money growth. Specifically, later observations on *(DM, FEDV, UN)* would be used to estimate the coefficients of the *DM* relation, and these coefficients would then be applied to the 1950 values of the independent variables to obtain $\widehat{DM}$ for 1950. However, it should be noted that the manner in which later observations affect earlier values of $\widehat{DM}$ is solely through pinning down the estimates of the coefficients in the *DM* equation. If individuals have information about the money growth structure beyond that conveyed in prior observations (for example, from the experiences of other countries or on theoretical grounds), then the use of the overall sample period, 1941–73, may be reasonable even for the earlier dates.

A procedure that avoids the use of later observations to generate earlier predictions involves obtaining $\widehat{DM}_t$ from a regression in which the coefficients are estimated from data only up to date t − 1. In this approach there would be as many *DM* equations (each incorporating data up to t − 1) as there were predicted values, $\widehat{DM}_t$. In this context it would also be natural to consider the possibility of weighting the observations so that more recent information was counted more heavily in forming predictions.[8]

In my earlier study (1975), which is available on request, I devoted a good deal of space to estimations that based money growth predictions solely on prior observations. Since it turned out that the implications for the analysis of unemployment were minor, I have not included this discussion. For the present analysis I use the estimated values of *DM* from the 1941–73 regression, equation (2), to form a time-series of anticipated money growth $\widehat{DM}$. Unanticipated money growth, $DMR \equiv DM\text{-}\widehat{DM}$, then corresponds to the residuals from this equation. The values of $\widehat{DM}$ and *DMR* from equation (2) are indicated in Table 1.[9]

II. Analysis of Unemployment

A. *Setup of the Equation*

The effects of monetary expansion on unemployment are measured by the impact of current and lagged values of unanticipated money growth, $DMR \equiv DM\text{-}\widehat{DM}$. The number of lags to introduce was not established from a priori reasoning, although Lucas (1975) presents a theoretical rationale for persistence effects of monetary shocks in this type of model. Empirically, it turned out that the current and two annual lag values of *DMR* had significant effects on unemployment.

Aside from monetary variables, the unemployment equation includes two "real" variables. The first is a measure of military conscription. The specific variable is[10]

$$MIL \equiv \frac{\text{Military personnel}}{\text{Male population aged 15–44}}$$

for years in which a "selective" military draft law was in effect (all years since 1946 except for April 1947 to June 1948 and 1970–73).[11] The

[8]Heavier weighting of recent observations can be rationalized along the lines of the adaptive regression model, as discussed in Thomas Cooley and Edward Prescott

[9]Note that $\widehat{DM}_t$ is calculated from the contemporaneous value of *FEDV*, rather than from a lagged value. The rationale is that the principal movements in *FEDV*, which are dominated by changes in wartime activity, would be perceived sufficiently rapidly to influence $\widehat{DM}$ without a lag. For example, in 1946 the value of $\widehat{DM}$ is much lower than in 1945 because of the contemporaneous downward movement in *FEDV*.

[10]Data sources are *Historical Statistics of the United States*, 1960, pp. 736, 8, and 10, and *Statistical Abstract of the United States*, various issues.

[11]A discussion of the draft law in the United States up to 1970 is contained in John Rafuse. The lottery draft period from 1970 to June 1973 (during which there were draft calls for 1970–71) was taken out since the lottery draft does not provide the same incentives to avoid unemployment as appear to operate during a selective draft. See the discussion below. Periods with zero draft calls, but with a selective draft law in effect (February 1949 to June 1950) were included with the draft period.

value $MIL = 0$ was entered for the nonselective draft law years. Values of MIL are tabulated from 1946 to 1975 in Table 2. Aside from the possible direct employment effect of conscription on draftees, a selective draft would provide incentives for eligible civilians to enter a low draft-probability status. One effect would involve the choice of remaining in school rather than entering the labor force—an effect that would reduce the measured unemployment rate if the affected individuals had an above-average tendency toward unemployment. A second effect involves the choice between working and unemployment for labor market participants. On this count, conscription would work toward reducing the unemployment rate if draft probabilities were highest, *ceteris paribus*, for unemployed persons.

The second real variable measures the impact of the minimum wage rate. This variable, tabulated under the heading $MINW$ from 1946 to 1975 in Table 2, is defined as the ratio of the applicable minimum wage to private, nonfarm average hourly earnings, multiplied by the proportion of covered nonsupervisory employment.[12] The $MINW$ variable would have a positive effect on the unemployment rate if the negative impact of the minimum wage on employment dominates the probable negative effect on labor force participation.[13]

The form for the systematic part of the unemployment equation,[14] used for annual observations over 1946–73, is[15]

$$(3)\quad log(U/(1-U))_t = a_1 DMR_t + a_2 DMR_{t-1} + a_3 DMR_{t-2} + a_4 MIL_t + a_5 MINW_t$$

Since the sample period begins in 1946, the values for DMR start in 1944. It may be worth noting that, since the dependent variable in equation (3) depends on a distributed lag of DMR values, the unemployment rate can be serially correlated even if the DMR (and the MIL and $MINW$ variables) were not. (The DMR values would not be serially correlated if $\widehat{DM}_t$ were an efficient predictor of DM_t, based on information that included an observation of DM_{t-1}.)

B. *Estimated Equation Based on Unanticipated Money Growth Rates*

With DMR measured as the residuals from equation (2), the estimated unemployment equation is, with standard errors in parentheses,[16]

[12]From 1947 to 1968 this variable was calculated by the Bureau of Labor Statistics and reported in Jacob Mincer, Table 1 6 For 1946 and 1969 to 1975 the variable is estimated from data contained in Armand Weiss, Tables 1–3

[13]I also considered an unemployment compensation variable, which was defined as average benefits per recipient relative to average hourly earnings multiplied by the fraction of covered employment In my initial investigations this variable was insignificant in the unemployment equation. However, I have recently recalculated the unemployment compensation variable to take account of taxes on earnings and to incorporate a fuller measure of unemployment compensation coverage This revised variable does turn out to have a significantly positive effect on the unemployment rate The other coefficients in the estimated unemployment equation are insensitive to the inclusion of this variable, except that the minimum wage variable becomes less important. I plan to report on these results more fully at a later time

[14]The form confines the unemployment rate to the interval (0, 1)

[15]Since DMR is based on estimated coefficients of the DM relation, equation (2), there would be small-sample problems of errors in the independent variables in equation (3). The main impact would seem to be a bias toward zero in the estimated DMR coefficients. In obtaining estimates of the a-coefficients in equation (3) and the α-coefficient in equation (1), it would be preferable to carry out a joint maximum likelihood estimation In contrast with my two-stage procedure, the choice of the α-estimates in the money growth equation would then give some weight to the effect on the fit of the unemployment equation, through the selection of the DMR values that enter into equation (3). In my procedure the α-estimates are chosen solely to obtain a least-squares fit in equation (1) For the case of normally distributed error terms, the a-estimates would be chosen in both cases to obtain a least-squares fit in equation (3), conditional on the DMR values. Since my procedure yields consistent estimates (assuming serially-independent error terms) and the alternative, non-linear procedure requires a large amount of numerical calculation, I have not carried out the joint maximum likelihood estimation

[16]I have carried out a similar analysis (1975) using the *log* of output (real GNP) instead of the unemployment rate as a dependent variable (with a time trend included as an additional explanatory variable). The results correspond in major respects to those discussed below for unemployment.

$$(4)\quad log(U/(1-U))_t = \underset{(.15)}{-3.07} - \underset{(2.1)}{5.8}DMR_t - \underset{(1.9)}{12.1}DMR_{t-1} - \underset{(1.9)}{4.2}DMR_{t-2} - \underset{(0.8)}{4.7}MIL_t + \underset{(.46)}{0.95}MINW_t,$$

$$R^2 = .78,\ \hat{\sigma} = .13,\ D.W. = 1.96,\ \text{average of } |U - \hat{U}| = .0043$$

Equation (4) includes a contemporaneous and two annual lag values of *DMR*. Additional lag terms were insignificant. The implied lag pattern (the form of which was not constrained *ex ante*) for unemployment behind unanticipated money growth has a triangular shape, with the strongest effect appearing after a one-year lag. The contemporaneous and two-year lag effects are of about equal size.[17] The t-values associated with a null hypothesis of a zero coefficient are 6.4 for DMR_{t-1}, 2.8 for DMR_t, and 2.2 for DMR_{t-2}. The F-value for the three *DMR* coefficients jointly is $F^3_{22} = 21.0$ (5 percent critical value = 3.1). More detailed aspects of the estimated *DMR* coefficients and of the estimated coefficients for the *MIL* and *MINW* variables will be discussed below.

In evaluating the fit of equation (4), a useful measure is the average absolute residual for implied estimates of the unemployment rate (generated from a straightforward, though not quite statistically valid, transformation of the dependent variable, $log(U/1 - U)$). This average value is .0043—that is, the average error in estimated unemployment rates is somewhat more than $\frac{4}{10}$ of a percentage point.[18]

The Durbin-Watson statistic from equation (4) of 1.96 indicates absence of first-order serial correlation in the residuals. This result is surprising, given the autocorrelated nature of the U-series,[19] since a lagged dependent variable was not included to soak up the serial correlation.[20] In fact, if $log(U/1 - U)_{t-1}$ is added to equation (4), its estimated coefficient is .09, standard error = .10, which differs insignificantly from zero.

C. *Results with Total Money Growth Rates*

Unemployment regressions have also been run based on total money growth rates *DM*, rather than on the unanticipated part of growth *DMR*. For a regression that includes a contemporaneous and two lagged values of *DM* along with the military and minimum wage variables, none of the estimated *DM* coefficients turns out to be individually significantly different from zero, and an F-test for joint significance yields the statistic $F^3_{22} = 2.6$, which is below the 5 percent critical value of 3.1. The fit of the regression is indicated by $R^2 = .38$—half that of the *DMR* equation. The inclusion of a third lag of *DM* into the unemployment equation has a negligible impact. When DM_{t-4} is included the fit improves noticeably, although the estimated coefficients on DM_{t-2} and DM_{t-3} are positive. The estimated equation with four lags is

$$(5)\quad log(U/(1-U))_t = \underset{(.34)}{-2.46} - \underset{(2.9)}{1.2}DM_t - \underset{(2.7)}{5.7}DM_{t-1} + \underset{(2.5)}{0.7}DM_{t-2} + \underset{(1.8)}{3.5}DM_{t-3} - \underset{(1.5)}{3.2}DM_{t-4} - \underset{(1.4)}{4.5}MIL_t - \underset{(1.0)}{0.3}MINW_t,$$

$$R^2 = .52,\ \hat{\sigma} = .20,\ D.W. = 1.68,\ \text{average of } |U - \hat{U}| = .0059$$

The F-value for the joint hypothesis that all five *DM* coefficients are zero is $F^5_{20} = 3.0$, which is

[17] The estimated coefficient of DMR_t could be biased toward zero if there is a contemporaneous policy feedback from U_t (current period unemployment) to $\hat{DM}_t$. The response of money growth to lagged unemployment was already taken into account in forming the anticipated money growth rate $\hat{DM}_t$. Presumably, this problem of within-period policy response would be lessened if the length of the observation period were reduced by moving to quarterly data. I plan to carry out that extension at a later time.

[18] I have redone the unemployment analysis with two alternative definitions of the money stock, M_2 and high-powered money (see the author, 1975, for details). It turns out that the M_1 definition is superior in terms of the fit for unemployment. In a form parallel to equation (4), the M_2 definition yields an R^2 of .31 with an average absolute error for U of .0076—about twice that of the M_1 form. For high-powered money the R^2 is .49 with an average absolute error for U of .0066.

[19] An autoregression of UN_t on UN_{t-1} yields the estimated coefficient .40, standard error = .14.

[20] Except for the indirect effect of lagged U on $\hat{DM}$. However, that effect was estimated from a separate equation, so that the usual problem of correlation between a lagged dependent variable and a serially-correlated error term would not arise here.

above the 5 percent critical value of 2.7. The fit of the equation with four lagged values of DM is indicated by $R^2 = .52$, average absolute error for $U = .0059$. Hence, the fit is still considerably poorer than that obtained in equation (4) with two lagged values of the DMR variable.

D. *Tests that Only Unanticipated Money Growth Affects Unemployment*

A key hypothesis of this study is that only the unanticipated part of money growth influences unemployment. This hypothesis can be tested by running a regression that includes simultaneously sets of DMR and DM variables, and then seeing whether the deletion of the DM variables, which amounts to a set of linear restrictions on the coefficients, produces a significant worsening of the fit. The resulting test statistic is $F^3_{19} = 1.4$ (5 percent critical value = 3.1) when two lagged values of DMR and DM are included, and $F^5_{15} = 2.0$ (5 percent critical value = 2.9) when four lagged values of each are included. Hence, the hypothesis that only the unanticipated part of money growth is relevant to unemployment is accepted by these tests.

The procedure can also be carried out in reverse by deleting the DMR values while retaining the DM values. When two lagged values of DMR and DM are included, the test statistic is $F^3_{19} = 15.7$. In the four-lag case the result is $F^5_{15} = 8.2$. Therefore, the reverse hypothesis that the DMR values are irrelevant to unemployment, given the DM values, can easily be rejected.

A point to stress about these tests is that they can be carried out at all only because predictors of DM_t other than its own history—DM_{t-1}, DM_{t-2}, etc.—have been included in the money growth equation. For example, suppose that $\widehat{DM}_t$ were generated solely as a function of DM_{t-1} say, $\widehat{DM}_t = \alpha_0 + \alpha_1 DM_{t-1}$. In this case a regression of unemployment on a series of DMR ($\equiv DM - \widehat{DM}$) values could not possibly fit better than a regression of the same form on a series of DM values that included one additional lagged term. The use of the DMR values would amount, in this situation, solely to imposing a restriction on the coefficients that describe the effect of the DM variables on unemployment, so that (if no adjustment is made for the difference in degrees of freedom) the DMR regression would necessarily show a poorer fit. Hence, the superior fit of the DMR form of the unemployment equation reflects the impact of the additional predictors—namely, the federal expenditure and lagged unemployment variables—that were included in the money growth equation.

To make this point directly I have obtained DMR values from money growth equations that involve solely the history of DM. An illustrative case, which includes 3 lagged values of DM over the 1941 to 1973 period, is the following:

$$(6)\qquad \underset{(.008)}{DM_t = .011} + \underset{(.17)}{.76DM_{t-1}} + \underset{(.21)}{.30DM_{t-2}} - \underset{(.14)}{.30DM_{t-3}},$$

$$R^2 = .77,\ \hat{\sigma} = .031,\ D.W. = 2.16$$

Calculating DMR values as the residuals from the above equation leads to the estimated unemployment equation for 1946–73,

$$(7)\qquad log(U/1-U)_t = \underset{(.31)}{-3.00} + \underset{(2.9)}{1.2DMR_t} - \underset{(2.7)}{4.9DMR_{t-1}} - \underset{(2.4)}{1.9DMR_{t-2}} - \underset{(1.3)}{2.6MIL_t} + \underset{(0.9)}{0.2MINW_t},$$

$$R^2 = .31,\ \hat{\sigma} = .23,\ D.W. = 0.95$$

which shows a substantially poorer fit than that obtained with the alternative DMR values from equation (2). Hence, a "naive" model that bases $\widehat{DM}_t$ solely on the history of money growth would be inadequate for explaining unemployment.[21]

Further perspective on the distinction between actual and unanticipated money growth can be obtained by substituting into the estimated unemployment relation, equation (4), from the condition $DMR_t \equiv DM_t - \widehat{DM}_t$, where $\widehat{DM}_t$ is generated from the estimated money growth relation, equation (2). The resulting "reduced

[21] If lagged values up to DM_{t-10} are included, the R^2 of the DM equation rises to .89 and that of the unemployment equation rises to .35. Allowing for first-order serial correlation of the error term in the DM equation does not materially affect any of these results.

form" expresses unemployment as a function of ($DM_t, \ldots, DM_{t-4}$); ($FEDV_t, \ldots, FEDV_{t-2}$); ($UN_{t-1}, \ldots, UN_{t-3}$); MIL_t; and $MINW_t$. Specifically, the coefficients that derive from this substitution are indicated as hypothesized values of the first column of Table 3.

TABLE 3—HYPOTHESIZED AND ESTIMATED COEFFICIENTS OF REDUCED FORM FOR UNEMPLOYMENT

	Hypothesized	Estimated	Standard Error
C	−1.2	−1.1	(0 5)
DM_t	−5.8	−2 5	(2 6)
DM_{t-1}	−10.7	−12.5	(2.9)
DM_{t-2}	0 8	−5.8	(4 6)
DM_{t-3}	5.2	4 4	(1 8)
DM_{t-4}	1 5	2 3	(1 6)
$FEDV_t$	0 5	0 3	(0.7)
$FEDV_{t-1}$	1.0	1.3	(0.4)
$FEDV_{t-2}$	0.3	0 8	(0.5)
UN_{t-1}	0 2	−0 3	(0.4)
UN_{t-2}	0.3	0 5	(0.2)
UN_{t-3}	0.1	0 2	(0.2)
MIL_t	−4.7	−8 8	(2 2)
$MINW_t$	0 9	−0.6	(0.7)

It is also possible to estimate the reduced form for unemployment in a direct, unconstrained manner—a process that yields the estimated coefficients and standard errors that are also shown in Table 3.

The use of the *DMR* form of the unemployment relation, equation (3), corresponds to a set of constraints on the manner in which the reduced form independent variables influence unemployment. Specifically, the use of equation (3) with *DMR* values generated from equation (2) amounts to reducing the number of independent coefficients to be estimated in the unemployment relation from (14) in the unconstrained reduced form to (6) in the *DMR* form.[22] If the *DMR* specification in equation (3) is appropriate, then these 8 coefficient constraints should not significantly worsen the fit of the unemployment equation—heuristically, the hypothesized coefficients in Table 3 should not differ "too much" from the estimated ones (taking account of standard errors). An overall test of the hypothesis is based on a comparison of restricted and unrestricted sums of squared residuals which leads to the statistic $F^8_{14} = 1.4$, which is less than the 5 percent critical value of 2.7. Hence, this test also supports the use of the *DMR* form of the unemployment equation.

The listing of the reduced form coefficients in Table 3 brings out another point, which relates to the discussion of observational equivalence in Sargent (1976). Namely, the *DMR* form of the unemployment equation is equivalent to a form that contains *DM* values (in this case up to DM_{t-4}), along with the *FEDV* and lagged *UN* variables (up to $FEDV_{t-2}$ and UN_{t-3}, respectively) that were included in the *DM* relation. The exclusion of the *FEDV* and lagged *UN* variables from the form of the unemployment relation, equation (3), constitutes a set of identifying restrictions that permits an observational separation between the *DMR* and *DM* forms of the unemployment equation. The above tests of the distinction between these two forms then amount to tests of the joint hypothesis that (a) $\widehat{DM}$ is generated in accordance with equation (2); (b) *DM* influences unemployment only in the form, $DMR \equiv DM - \widehat{DM}$; and (c) the *FEDV* and lagged *UN* variables that appear in equation (2) do not enter directly in equation (3). Of course, the acceptance of the joint null hypothesis by the above statistical tests provides support for each element of the hypothesis, namely for (a) and (b), which were the main objects of interest.

It would be possible, nevertheless, to interpret the estimated reduced form for unemployment (Table 3, col. 2) as indicating the influence of actual money growth *DM*, along with direct influences of the *FEDV*, lagged *UN*, *MIL*, and *MINW* variables (with the coefficients of the *DM*, *FEDV*, and lagged *UN* variables satisfying the restrictions implied by the *DMR* form out of pure coincidence). However, this interpretation leaves a number of results that require a theoretical explanation: 1) the *positive* effect of the

[22]There are also 5 coefficients to be estimated in the *DM* equation, but this estimation was carried out separately from the fitting of the unemployment relation.

FEDV variables on unemployment, in contrast with the negative effect that would be predicted along Keynesian lines; 2) the presence of positive coefficients on DM_{t-3} and DM_{t-4}; and 3) the stronger (positive) contribution of UN_{t-2} than of UN_{t-1}. These three sets of results are readily explained by the theory that relates unemployment to *DMR* values.

E. *Properties of the Estimated Unemployment Equation*

I will now discuss some detailed properties of the estimated *DMR* form of the unemployment relation, which is rewritten here for convenience,

$$(4)\quad log(U/1-U)_t = \underset{(.15)}{-3.07} - \underset{(2.1)}{5.8}DMR_t - \underset{(1.9)}{12.1}DMR_{t-1} - \underset{(1.9)}{4.2}DMR_{t-2} - \underset{(0.8)}{4.7}MIL_t + \underset{(.46)}{0.95}MINW_t$$

Consider the magnitudes of the estimated *DMR* coefficients. The coefficient of −12 on DMR_{t-1} implies that an increase by 1 percentage point per year in the unanticipated money growth rate would reduce next year's unemployment rate by a proportion of about 12 percent, or by about $\frac{6}{10}$ of a percentage point at an initial unemployment rate of 5 percent. However, the contemporaneous impact of this *DMR* shift would be only about half as large. If the increase in *DMR* by 1 percentage point per year were sustained over a three-year period (which would be an unusual event), then the full effect would be a reduction of the unemployment rate by about 1 percentage point.

It should be stressed that the lag pattern for money growth that is described in equation (4) refers to unanticipated rather than actual money growth. The implied lag pattern in terms of actual money growth—given the money growth relation as estimated in equation (2)—is shown as the hypothesized coefficients in Table 3. Because of the positive effects of DM_{t-1} and DM_{t-2} on the current value of anticipated money growth, the lag pattern for unemployment in terms of *DM* differs markedly from that in terms of *DMR*. Two important differences are, first, the "mean" lag effect from *DM* to unemployment is shorter than that associated with *DMR*; and, second, there are positive coefficients in the *DM* form even when the *DMR* form is restricted to negative coefficients. Quantitatively, the lag pattern for *DM* that is shown in column 1 of Table 3 accords with the well-known 6- to 18-month lag between (actual) money growth and economic activity that has been reported by Milton Friedman (p. 180). A lack of distinction between actual and unanticipated money growth can also account for some of the apparent variability of the lag in Friedman's results (pp. 180–81).

Equation (4) also indicates the importance of the military variable (t-value of 5.9). The magnitude of the effect implied by the coefficient of −4.7 is that an increase by 1 percentage point in the ratio of military personnel to the male population aged 15–44 would reduce the unemployment rate by a proportion of about 4.7 percent; that is, by about $\frac{2}{10}$ of a percentage point at $U = 5$ percent. Expressed alternatively, if changes in the labor force are neglected, an increase by an amount X in the number of military personnel would reduce the number of unemployed by about $0.5X$ (assuming that $U = .05$ and that the ratio of the labor force to the male population aged 15–44 takes on its 1973 value of 2.0).[23]

The estimated minimum wage coefficient in equation (4) is positive and has a t-value of 2.1. Using the 1973 values of average hourly earnings ($3.92) and fraction covered (.79), and

[23]If the distinction between selective draft and nonselective draft years is dropped (which affects 1947–48 and 1970–73), the estimated unemployment rate equation becomes

$$log(U/1-U)_t = \underset{(.20)}{-2.88} - \underset{(2.5)}{4.5}DMR_t - \underset{(2.3)}{10.1}DMR_{t-1} - \underset{(2.2)}{1.5}DMR_{t-2} - \underset{(1.6)}{7.1}MIL_t + \underset{(.56)}{1.11}MINW_t,\ R^2 = .68,$$

$$\hat{\sigma} = .16,\ D.W. = 1.38,\ \text{average of } |U-\hat{U}| = .0049$$

The fit of the equation is poorer than that of equation (4), but the general implications are not altered.

starting from $U = .05$, the implication is that an increase by $1 in the minimum wage would raise the unemployment rate by about 1 percentage point. Viewed alternatively, if the minimum wage ($1.60 in 1973) were set to zero, the estimated fall in the unemployment rate would be by about 1⅓ percentage points.

Given the estimated relation from equation (4), it is possible to calculate values of unemployment associated with $DMR = 0$ for all t—that is, with fully anticipated current and past monetary expansion. I will refer to these unemployment rates as natural values *(UNAT)*.[24] In the present setup, the natural unemployment rate depends on the values of the military and minimum wage variables and on the constant term. Values of *UNAT* derived from equation (4), and the values of *MIL* and *MINW* shown in Table 2, are indicated from 1946 to 1975 in Table 1. This table also contains actual unemployment rates and the estimated values and residuals from equation (4). The pattern of results in this table is as follows.

With the end of World War II and the associated drop in military personnel, the estimated natural unemployment rate rose from about 1.5 percent to about 3.5 percent in 1946 and 5 percent in 1947–48 (partially non-draft law years). Although there was a large cutback in money growth, from rates above 15 percent per year during World War II to 6.8 percent in 1946 and 4.7 percent in 1947, the money growth equation implies that this cutback was anticipated because of the sharp decline in federal expenditure. In fact, the estimated values, $\widehat{DM} = 5.5$ percent in 1946 and 3.8 percent in 1947, imply that these two years were characterized by unanticipated monetary expansion. Accordingly, the unemployment rates for 1946–48 remained at about 4 percent—below the natural rate for 1947–48. The unanticipated monetary contraction of 1948–49 ($DMR = -.012$ and $-.023$, respectively) implied increases in the unemployment rates for 1949 and 1950.

For the Korean War years of 1951–53, an expansionary element was an increase in the military variable that lowered the natural unemployment rate to 3.5 to 4 percent. This factor, combined with unanticipated monetary expansion from 1950 to 1952 (*DMR* values of .019, .018, and .012, respectively), led to unemployment rates in the neighborhood of 3 percent for 1951–53. From the end of the Korean War through 1969, the maintenance of a selective draft law with high levels of military personnel implied small variations in the natural unemployment rate. In particular, with *UNAT* confined to a range of 4.0 to 4.4 percent from 1956 to 1969, movements in the natural rate have a minor effect on estimated unemployment rates during this period.

In 1954 the unanticipated monetary contraction of 1953 ($-.017$) was the main contributor to the rise in unemployment (though my U-estimate of .044 is below the actual value of .052). For 1954–55, the unanticipated parts of money growth were small, implying values of U near the natural rate of 4 percent for 1955–56. The unanticipated monetary contraction in 1956 ($-.009$) led to an estimated U-value for 1957 of 4.9 percent, although the actual value was only 4.1 percent. On the other hand, my estimate for U in 1958 is 5.4 percent (reflecting the additional monetary contraction of $-.018$ in 1957), which substantially underestimates the actual value of 6.5 percent. For 1959-60, the estimates are about ½ percentage point below the actual values, which were themselves about 1 percentage point above the natural rates.

Perhaps the most interesting monetary behavior of the post-World War II period is the absolute contraction of money that occurred during 1960. This behavior represented the first absolute decline in money since 1949, but more significantly, the estimate for anticipated money growth in 1960 is 3.0 percent, as contrasted with 1.3 percent for 1949. Hence, the unanticipated monetary contraction for 1960 was -3.1 per

[24] Because of nonlinearities, these values differ from expected unemployment rates derived from equation (4) with an additive, constant variance error term. The (positive) gap between the expected unemployment rate and the natural rate, as defined, increases with the variance of the error term and with the variance of *DMR*.

cent—the largest absolute value of *DMR* for the entire post-World War II period. According to the estimated equation, this large negative value of *DMR* for 1960 accounted for the sharp rise in the unemployment rate in 1961 to over 6 percent—about 2 percentage points above the natural rate.

From 1963 to 1967 there was a period of monetary stability, in the sense of small deviations between actual and anticipated values of *DM*. The response in *U* was a gradual downward movement, first to the natural rate in 1965, and then slightly below in 1966–67. There was then a sharp monetary expansion in 1968 (*DMR* = .026), which ended the brief period of "constant growth rate rule" for money.[25] In 1968–69 the unemployment rate of 3.4 percent was about 1 percentage point below the natural rate.

The explanation of behavior in 1970 is complicated since it hinges on the treatment of the switch to the lottery draft as equivalent, in terms of unemployment effects, to a removal of conscription (see fn. 23). The assumption that the military variable was zero from 1970 on implies a natural rate since 1970 of 6 to 6.5 percent (depending on the value of the *MINW* variable)—an increase of $1\frac{1}{2}$ to 2 percentage points from the 4.4 percent value for 1969. Given the rise in the natural rate, the maintenance of 1970 unemployment at only 4.7 percent of the labor force reflected the continuing impact of the strong monetary expansion that occurred in 1968–69. The monetary behavior from 1971 to 1973 was expansionary, and the unemployment rate remained $\frac{1}{2}$ to 1 percentage point below the natural rate during this period.

F. *Unemployment Predictions*

The unemployment and money growth rate relations, estimated from data up to 1973, can be used to form projections for 1974 and beyond. For 1974, the predicted value of *DM* is 5.6 percent per year, as compared to an actual value of 5.5 percent. Hence, the *DMR* value for 1974 is close to zero. The prediction from equation (4) for the 1974 unemployment rate is 5.6 percent, which almost coincides with the actual value of 5.5 percent.

For 1975, the predicted value of *DM* (conditioned on the value DM_{t-1} = .055 for 1974) is 6.2 percent per year. Since the actual value of *DM* for 1975 is 4.2 percent, the monetary contraction during this year is measured by *DMR* = −2.0 percent. Using the *ex post* values, *DMR* = −.001 for 1974 and −.020 for 1975, the "predicted" value for 1975 unemployment turns out to be 7.1 percent. Since the actual average of unemployment rates during 1975 is 8.3 percent, there is an underprediction of unemployment by about the same magnitude as for the 1958 contraction.

For 1976 and 1977 (using the value, *FEDV* = .18, which is the average over the 1960–75 period), the predicted value for *DM* is 6.5 percent per year.[26] Using values of *DMR* = 0 for 1976 and beyond (which is appropriate *ex ante*), assuming a zero value for *MIL*, and using the values of *MINW* that are shown in Table 2, the predicted unemployment rates are 8.1 percent for 1976, 6.8 percent for 1977, and 6.1 percent (the natural rate) for 1978 and beyond. Based on observations for the first few months, it appears that the model will overpredict 1976 unemployment.

III. Some Policy Implications

Acceptance of the hypothesis that only the unanticipated part of money growth affects unemployment has some important policy implications. One result is that the systematic feedback from unemployment to money growth that appears in equation (2) has no implications for

[25] I use this expression to signify predictability of *DM*, rather than constancy per se

[26] This high value of $\hat{DM}$ reflects the high value of lagged unemployment. It may be preferable to measure unemployment relative to its perceived long-run value, which has apparently increased since 1970 (see fn. 7). This modification would lower the values of $\hat{DM}$ for the 1970's, but a quantitative adjustment would require a measure of the perceived long-run value of unemployment.

the time path of unemployment itself—a result that accords with the theoretical propositions in Sargent and Wallace and the author (1976a). Only movements in money that depart from the usual countercyclical response affect subsequent unemployment rates.[27] This observation raises questions concerning the rationality of the countercyclical policy response that appears in equation (2). One possibility is that the reaction of money to lagged unemployment reflects optimal public finance considerations (see Section I and the author, 1976b), rather than an attempt at economic stabilization.

Similar conclusions apply to the response of money to the federal budget variable, *FEDV*. Increases in federal expenditure above its normal level (with the military variable held fixed) reduce unemployment only if the accompanying increase in money is larger than the usual amount.[28] In fact, if actual money growth is held constant, an increase in *FEDV* raises unemployment because of the associated increase in anticipated money growth. (Some preliminary results indicate that this effect is important during the middle 1930's.)

IV. Conclusions and Extensions

The starting point for this study was the hypothesis that only unanticipated movements in money would affect economic activity. That hypothesis was quantified by interpreting anticipated money growth as the amount that could have been predicted based on the historical relation between money growth and a specified set of explanatory variables. For the United States from 1941 to 1973 these variables included a measure of federal expenditure relative to normal, a lagged unemployment rate, and two annual lag values of money growth. Unanticipated money growth was then measured as actual growth less the amount obtained from this predictive relationship. The current and two annual lag values of unanticipated money growth were shown to have considerable explanatory value for unemployment. Further, some statistical tests confirmed the underlying hypothesis that actual money growth was irrelevant for unemployment, given the values of unanticipated money growth.

The results reported in this paper would be more reliable if they could be replicated for other experiences. For the United States I am currently working on the unemployment and output experiences back to 1890. Since the structure of the money growth process prior to World War II appears different from that estimated for the 1941–73 period, the long-period evidence will permit a much more powerful test of the hypothesis that only unanticipated money growth affects unemployment. Further, it will be possible to test the hypothesis advanced by Lucas (1973) that shifts in the prediction variance of money would alter the response of unemployment to monetary shocks.

Finally, although the present analysis was directed toward the effects of money on unemployment (with related implications for output), the division of money growth into anticipated and unanticipated parts also has important implications for inflation. I plan to deal with this topic in a subsequent paper

[27]However, the present analysis has not dealt with the possible temporary impact of structural shifts in the money growth process, as discussed theoretically in John Taylor Such shifts did not appear to be important over the 1941–73 period (see the author, 1975)

[28]I attempted to find a direct fiscal effect on unemployment by entering the full-employment federal government deficit (measured as a ratio to the outstanding stock of privately held public debt) into the unemployment equation This variable was insignificant (estimated coefficient of −0.7, standard error = 1.0), as were lagged values of the deficit and measures of the deficit relative to its "anticipated" value

REFERENCES

R. J. Barro, "Unanticipated Money Growth and Unemployment in the United States," work. pap., Univ. Rochester, July 1975.

———, (1976a) "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

———, (1976b) "Optimal Revenue Collection and the Money Growth Rate," unpublished, 1976.

P. Cagan, "The Monetary Dynamics of Hyperinflation," in Milton Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

T. F. Cooley and E. Prescott, "Varying Parameter Regression: A Theory and Some Applications," *Annals Econ. Soc. Measurement*, 1973, *2*, 463–74.

M. R. Darby, "Three-and-a-half Million U.S. Employees Have Been Mislaid; or, an Explanation of Unemployment, 1934–1941," *J. Polit. Econ.*, Feb. 1976, *84*, 1–17.

M. Friedman, "The Supply of Money and Changes in Prices and Output," in his *The Optimum Quantity of Money and Other Essays*, Chicago 1969.

R. E. Lucas, "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

——, "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

——, "An Equilibrium Model of the Business Cycle," *J. Polit. Econ.*, Dec. 1975, *83*, 1113–44.

J. Mincer, "Unemployment Effects of Minimum Wages," *J. Polit. Econ.*, Aug. 1976, Part 2, *84*, S87–S104.

J. F. Muth, "Optimal Properties of Exponentially Weighted Forecasts," *J. Amer. Statist. Assn.*, June 1960, *55*, 299–306.

E. S. Phelps, "Inflation in the Theory of Public Finance," *Swedish J. Econ.*, Mar. 1973, *75*, 67–82.

J. L. Rafuse, "United States' Experience with Volunteer and Conscript Forces," *Studies Prepared for the President's Commission on an All-Volunteer Armed Force*, Vol. II, Washington 1970.

T. J. Sargent, "A Note on the 'Accelerationist' Controversy," *J. Money, Credit, Banking*, Aug. 1971, *3*, 721–25.

——, "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," *J. Polit. Econ.*, June 1976, *84*, 631–40.

T. J. Sargent and N. Wallace, "'Rational' Expectations, the Optimal Monetary Instrument and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

J. B. Taylor, "Monetary Policy during a Transition to Rational Expectations," *J. Polit. Econ.*, Oct. 1975, *83*, 1009–21.

A. Weiss, Statement before the Subcommittee on Labor Standards of the House Education and Labor Committee, Nov. 1975.

U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1957*, Washington 1960.

——, *Statistical Abstract of the United States*, Washington, various issues.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, various years.

Mean-Risk Analysis with Risk Associated with Below-Target Returns

By PETER C. FISHBURN*

The general concern of this paper is models for choice among mutually exclusive investment opportunities or portfolios having uncertain returns. This group includes a variety of parametric models based on means, variances, semivariances, loss probabilities, etc., and expected utility models, among which I include the stochastic dominance models. Although these models have many similarities, their differences—especially in computational efficiencies and implications about decision agents' preferences—provide a continuing source of controversy and impetus for research.

The specific concern of the paper is a class of models that offers at least a modest degree of reconciliation among the different viewpoints involved in this controversy. A representative of this class, referred to as an α-t model, appears to have reasonable computational possibilities—although not perhaps to the extent achieved for mean-variance analysis—along with a fair degree of compatibility with the primary concerns expressed by investment managers, with von Neumann-Morgenstern utility functions for investment decisions that have appeared in the literature, and with stochastic dominance relationships. As noted below and in the next section, specific forms of the α-t model have been examined in some detail by others.

In my basic presentation, no distinction will be made between returns expressed in monetary units and returns expressed as percentages. Uncertainty in a portfolio's return is assumed to be expressed by a probability distribution function F, with $F(x)$ the probability of getting a return not exceeding x and $F(x^-) = sup\ \{F(y)\colon y < x\}$ the probability of doing worse than x. With no loss in reality, each F is assumed to be bounded in the sense that $F(x) = 0$ and $F(y) = 1$ for some real x and y. The mean and variance of F will be denoted by $\mu(F)$ and $\sigma^2(F)$, respectively. Since the boundedness assumption can be removed only at the expense of more involved mathematics, it will be maintained through the paper.

The general model that I shall consider is a mean-risk dominance model in which risk is measured by a probability-weighted function of deviations below a specified target return t. The special case of the general model mentioned above is the α-t model in which risk is defined by the two-parameter function

$$(1) \qquad F_\alpha(t) = \int_{-\infty}^{t} (t-x)^\alpha \, dF(x) \qquad \alpha > 0$$

This integral and integrals written later are Lebesgue-Stieltjes integrals. Fixed values of α and t are used in (1) for all distributions in a particular situation. A special case of the α-t model is the mean-target semivariance model mentioned by Harry Markowitz (1959) and discussed in detail by James Mao (1970a, b), William Hogan and James Warren (1972, 1974), and Burr Porter. These authors show that the mean-target semivariance model, for which $\alpha = 2$, has several attractive features including a close correspondence with criteria of choice reported by investment managers and an intimate connection with second degree stochastic dominance (see Porter). However, there is no compelling a priori reason for taking $\alpha = 2$, and I shall argue that different values of α can approximate a wide variety of attitudes towards the risk of falling below the target return. Moreover, it will be shown that many of the attractive features of the mean-target semivariance model are shared by the more flexible α-t model. Thus

*Research professor of management science, College of Business Administration, The Pennsylvania State University. This research was supported by the Office of Naval Research

the general spirit of this paper is similar to that which underlies the analyses of Mao, Hogan and Warren, and Porter. Its main divergence from these prior works is its concern with a more general characterization of risk. Since this generality is achieved by (1) without increasing the number of parameters used in the target semivariance measure, the α-t model offers greater applicability in mean-risk dominance with no increase in the basic complexity of the model.

The paper is organized as follows. The general model is first placed in perspective with other parametric dominance models, and a few reasons for its selection as a potentially useful mean-risk dominance model are noted along with implications of the α-t model and the estimation of α. Congruence of the model with the expected utility theory of John von Neumann and Oskar Morgenstern is then discussed. The form of utility function that is consistent with the associated mean-risk tradeoff correspondent of our general dominance model is specified, and the literature of expected utility in investment contexts is reviewed to determine the adequacy of the α-t model to reflect risk attitudes revealed therein. The technical discussion concludes by showing how the efficient sets obtained with different values of α relate to efficient sets which result from first, second, and third degree stochastic dominance analysis. A summary of the results is given at the end of the paper.

I. The General Model

With the exception of the stochastic dominance models, which are discussed later in the paper, most dominance models in the investment and capital budgeting literature declare that distribution F is better than distribution G whenever $\mu(F) \geq \mu(G)$ and $r(F) \leq r(G)$ with at least one strict inequality, where r is a real-valued risk function defined on the distributions The familiar mean-variance or E-V dominance model developed by Markowitz (1952, 1959), James Tobin (1958, 1965), William Sharpe (1963, 1964), and John Lintner (1965a, b), among others, has $r = \sigma^2$ or $r = \sigma$. William Baumol's variation of this model has $r(F) = K\sigma(F) - \mu(F)$ with $K > 0$. Suppose e is the point of no gain and no loss. Then, when risk is taken as the probability of loss (Markowitz, 1959; Dean Pruitt), $r(F) = F(e^-)$; and, when risk is taken as weighted losses (Evsey Domar and Richard Musgrave), $r(F) = \int_{-\infty}^{e} (e - x)\,dF(x)$.

Most of these risk measures, as well as others, are special cases of Bernell Stone's generalized risk measure which has

$$(2) \quad r(F) = \int_{-\infty}^{\gamma(F)} |x - \eta(F)|^{\alpha}\, dF(x) \qquad \alpha \geq 0$$

where $\eta(F)$ is a reference level of wealth from which deviations are measured, α is a measure of the relative impact of large and small deviations, and $\gamma(F)$ is the range parameter that specifies what deviations are to be included in the risk measure. Equation (2) gives the below-mean semivariance measure of risk when $\alpha = 2$ and $\gamma(F) = \eta(F) = \mu(F)$. The target semivariance is obtained with $\alpha = 2$ and $\gamma(F) = \eta(F) = t$.

I shall presume that the reader is familiar with at least several of the limitations and criticisms of mean-variance analyses set forth, for example, by Markowitz (1959), James Quirk and Rubin Saposnik, Karl Borch (1963, 1968, 1969, 1974), G. O. Bierwag and M. A. Grove, Paul Samuelson (1967, 1970), Martin Feldstein, Clayton Alderfer and Harold Bierman, S. C. Tsiang (1972, 1974), John Chipman, Alvin Klevorick, Bierwag, Haim Levy, Samuelson and Robert Merton, and the author (1975), the general trend of which suggests that mean-variance analysis results should not be taken very seriously unless the probability distributions used in the analysis satisfy certain restrictions. However, even when such restrictions are satisfied, or approximately satisfied, there is a contention, set forth by Domar and Musgrave, Markowitz (1959), and Mao (1970a, b), among others, that decision makers in investment contexts very frequently associate risk with failure to attain a target return. To the extent that this contention is correct, it casts serious doubt on variance—or, for that matter, on any

measure of dispersion taken with respect to a parameter (for example, mean) which changes from distribution to distribution—as a suitable measure of risk.

The idea of a mean-risk dominance model in which risk is measured by probability-weighted dispersions below a target seems rather appealing since it recognizes the desire to come out well in the long run while avoiding potentially disastrous setbacks or embarrassing failures to perform up to standard in the short run. The general model of this type can be expressed by

(3) F dominates G if and only if $\mu(F) \geq \mu(G)$ and $\rho(F) \leq \rho(G)$ with at least one strict inequality,

where $\rho(F)$ is defined by

$$\rho(F) = \int_{-\infty}^{t} \varphi(t - x)\, dF(x) \tag{4}$$

in which $\varphi(y)$ for $y \geq 0$ is a nonnegative nondecreasing function in y with $\varphi(0) = 0$ that expresses the "riskiness" of getting a return that is y units below the target Although this general model is of interest in itself, I shall focus on its α-t form since this form is much simpler and requires estimation of only a single parameter α instead of a general φ function. However, some of my later comments will be addressed to the general model of (3) and (4).

With $F_\alpha(t)$ defined by (1), the specialization of (3) and (4) which gives the α-t model will be written as

(5) $F\,P(\alpha, t)\,G$ if and only if $\mu(F) \geq \mu(G)$ and $F_\alpha(t) \leq G_\alpha(t)$ with at least one strict inequality,

where $P(\alpha, t)$ denotes the specified dominance relation and t and α are real numbers with $\alpha > 0$. In this form the risk measure is a special case of (2). For continuity purposes, (5) is extended to $\alpha = 0$ by defining

$$F_0(t) = \lim_{\alpha \to 0} F_\alpha(t)$$

so that $F_0(t) = F(t^-)$, the probability of getting less than t. In addition, $P(\infty, t)$ can be defined by taking

$F_\infty(t) \leq G_\infty(t)$ if and only if there is a number N such that $F_\alpha(t) \leq G_\alpha(t)$ for all $\alpha \geq N$

$F_\infty(t) < G_\infty(t)$ if and only if there is a number N such that $F_\alpha(t) < G_\alpha(t)$ for all $\alpha \geq N$

Letting $x_0 = inf\ \{x\colon F(x) \neq G(x)\}$, it can be shown that $F_\infty(t) < G_\infty(t)$ if $x_0 < t$ and either $F(x_0) < G(x_0)$ or $\{F(x_0) = G(x_0)$ and there is a $\delta > 0$ such that $F(x) < G(x)$ for all x strictly between x_0 and $x_0 + \delta\}$. In particular, if $inf\ \{x\colon G(x) > 0\} < inf\ \{x\colon F(x) > 0\}$ and $inf\ \{x\colon G(x) > 0\} < t$, then $F_\infty(t) < G_\infty(t)$, so that $P(\infty, t)$ is partially governed by the worst possible returns.

In addition to the extreme cases associated with probability of failing to meet the target return ($\alpha = 0$) and with a comparison of worst possible outcomes ($\alpha = \infty$), the α-t model includes the target semivariance measure of risk ($\alpha = 2$) and risk as given by the conditional expected linear deviation below target times the probability of falling below the target ($\alpha = 1$).

A closer look at (5) reveals some interesting aspects of an α-t model in the following theorem.

THEOREM 1: *Suppose* (5) *holds with* $\alpha > 0$. *Then*:

(a) *If* $F(t^-) = G(t^-) = 0$, *then* $F\,P(\alpha, t)\,G$ *if and only if* $\mu(F) > \mu(G)$;

(b) *If* $\mu(F) = \mu(G)$, $F(t^-) = 0$ *and* $G(t^-) > 0$, *then* $F\,P(\alpha, t)\,G$;

(c) *If F is a degenerate distribution or sure-thing that assigns probability* 1 *to* $t - d$ *with* $d > 0$, *and if G is a nondegenerate distribution that has* $G(t) = 1$ *and* $\mu(G) = \mu(F) = t - d$, *then* $F\,P(\alpha, t)\,G$ *if and only if* $\alpha > 1$, *and* $G\,P(\alpha, t)\,F$ *if and only if* $\alpha < 1$.

Parts (a) and (b) are obvious from (5) since $F_\alpha(t) = G_\alpha(t) = 0$ in part (a) and $F_\alpha(t) = 0 < G_\alpha(t)$ in part (b). Part (a) says that $P(\alpha, t)$ is completely determined by expected returns when all possible returns for F and G lie at or above the target t. This of course is at variance with the observation that many people are risk

averse in the sense that a certain return of $(x + y)/2$ will be preferred to a 50-50 gamble for x or y when $x \neq y$. However, the idea of using expected return as the decision criterion when all returns are as good as the specified target has some appeal in an investment or capital budgeting context. The general notion of aversion to risk shows up in part (b), which says that if two distributions have equal means and one is certain to give a return as good as t while the other has positive probability of yielding a return less than t, then the former dominates the latter. This illustrates the central role played by the target return in an α-t model.

Although a form of risk aversion appears in part (b) of the theorem, part (c) reveals that the model may have a risk-seeking or "gambling" aspect when $\alpha < 1$ and all returns for F and G are at or below the target. For example, if $\delta > 0$ and returns $x - \delta, x, x + \delta$ are all below t, then the 50-50 gamble for $x - \delta$ or $x + \delta$ dominates x when $\alpha < 1$. Thus $\alpha < 1$ characterizes an individual who is willing to gamble at fair odds in an attempt to minimize the extent to which his return falls short of t. On the other hand, if $\alpha > 1$ then x as a sure-thing will dominate the 50-50 gamble for $x - \delta$ or $x + \delta$. A proof of part (c) is given in the Appendix.

I conclude this section with a few remarks on the determination of the two parameters in the α-t model. Depending on context and the circumstances of the decision maker or his firm, t might be formulated as a ruinous return, as the zero profit return, as the return available from an insured safe investment, or as a target which reflects a general attitude towards acceptable performance in the firm.

Given t, α is supposed to reflect the decision maker's feelings about the relative consequences (personal, corporate, etc.) of falling short of t by various amounts. If his main concern is failure to meet the target without particular regard to the amount, then a small value of α is appropriate. On the other hand, if small deviations below target are relatively harmless when compared to large deviations, then a larger value of α is indicated. As indicated above by Theorem 1(c), $\alpha = 1$ is the point which separates risk-seeking from risk-averse behavior with regard to returns below target.

The obvious way to estimate α on the basis of (5) is to work with gambles or distributions with equal means and to presume that (5) reflects the decision maker's preferences at least to the extent that he prefers F to G whenever $\mu(F) = \mu(G)$ and $F_\alpha(t) < G_\alpha(t)$. Letting $d > 0$ be a "significant difference," Theorem 1(c) shows that $\alpha \leq 1$ if a 50-50 gamble for t or $t - 2d$ is preferred to $t - d$ as a sure-thing, and that $\alpha \geq 1$ if $t - d$ is preferred to the 50-50 gamble.

Given $\alpha \leq 1$, the sure-thing for $t - d$ is then compared to the gamble which gives $t - 2d$ with probability p and $t + d(2p - 1)/(1 - p)$ with probability $1 - p$, where $p \geq 1/2$. The sure-thing has mean $t - d$ and risk d^α; the gamble has mean $t - d$ and risk $p(2d)^\alpha$. If the gamble is preferred to the sure-thing then $p(2d)^\alpha < d^\alpha$, or $\alpha \leq \log(1/p)/\log 2$. If the sure-thing is preferred to the gamble then $\alpha \geq \log(1/p)/\log 2$. Hence if p_0 is a value of p at which the two are approximately indifferent, then α is approximately $\log(1/p_0)/\log 2$. Thus, if $p_0 = 2/3$ then $\alpha \approx .58$, and if the gamble is preferred to the sure-thing for all $p < 1$ then $\alpha = 0$ is indicated

Given $\alpha \geq 1$, it is not possible to estimate α more precisely using a sure-thing and a gamble whose mean equals the sure-thing, since the gamble will never be preferred to the sure-thing if (5) reflects preferences as indicated above. In this case we use two two-point gambles. The first, which we denote G_1, has probability $p \leq 1/4$ for $t - 2d$ and $1 - p$ for t, with $\mu(G_1) = t - 2pd$ and $G_{1\alpha}(t) = p(2d)^\alpha$. The second, denoted G_2, is a 50-50 gamble between $t - d$ and $t + d(1 - 4p)$, with $\mu(G_2) = t - 2pd = \mu(G_1)$ and $G_{2\alpha}(t) = d^\alpha/2$. If G_1 is preferred to G_2 then $p(2d)^\alpha \leq d^\alpha/2$, or $\alpha \leq \log(1/(2p))/\log 2$; if G_2 is preferred to G_1 then $\alpha \geq \log(1/(2p))/\log 2$. Hence if p_0 is a value of p at which G_1 and G_2 are approximately indifferent, then $\alpha \approx \log(1/(2p_0))/\log 2$. If G_2 is preferred to G_1 for all $p > 0$, then $\alpha = \infty$ is indicated.

Since the α-t model is designed to approximate a potentially much more complex state of affairs, it would not be too surprising to find that the estimate of α is sensitive to the value of d used in the procedure even though d does not appear in the estimate. For example, if G_1 in the preceding paragraph is approximately indifferent to G_2 for a given value of d, then (5) suggests that indifference should not be changed by changes in d. If the p_0 which gives indifference depends in a significant way on d, then the α-t model may be inappropriate for the situation at hand.

II. Congruence with Expected Utility

The dominance model of (3) and (4) is a relaxed version of the more precise preference model that presumes that a decision maker's preferences depend entirely on $\mu(F)$ and $\rho(F)$ for the relevant distributions and that these preferences are weakly (totally) ordered. Corresponding to this more precise preference model, we shall say that the decision maker's preferences satisfy a mean-risk utility model with risk defined by (4) if and only if there is a real-valued function U in mean and risk such that, for all relevant distributions F and G,

$$F > G \text{ if and only if } U(\mu(F), \rho(F)) > U(\mu(G), \rho(G)) \tag{6}$$

with U increasing in μ and decreasing in ρ. In (6), $>$ is the preference relation with $F > G$ read as "F is preferred to G."

It is entirely possible that a decision maker's preferences satisfy a mean-risk utility model without also satisfying the von Neumann and Morgenstern axioms for expected utility. That is, the validity of (6) does not depend on whether $>$ is consistent with a set of axioms that imply that there exists a real-valued function u on returns such that

$$F > G \text{ if and only if } \int_{-\infty}^{\infty} u(x)dF(x) > \int_{-\infty}^{\infty} u(x)dG(x) \tag{7}$$

However, since the expected utility model is considered by many writers as the best approach to rational decision making in the uncertainty setting, we shall specify what must be true of u so that the preference relations in (6) and (7) are identical.

In the following theorem, an "origin" and "scale unit" for u are specified in a way that simplifies the resulting expressions. Positive linear transformations of u, of the form $c_1 u + c_2$ with $c_1 > 0$, can be made without affecting the general conclusion. The theorem is proved in the Appendix.

THEOREM 2: *Suppose that, for all bounded distribution functions F and G, the mean-risk utility model with risk defined by (4) is congruent with the expected utility model in the sense that*

$$U(\mu(F), \rho(F)) > U(\mu(G), \rho(G)) \tag{8}$$

if and only if

$$\int_{-\infty}^{\infty} u(x)dF(x) > \int_{-\infty}^{\infty} u(x)dG(x)$$

with U increasing in μ, decreasing in ρ, and with $u(t) = t$ and $u(t+1) = t+1$. Then there is a positive constant k such that

$$u(x) = x \qquad \text{for all } x \geq t \tag{9}$$

$$u(x) = x - k\varphi(t-x) \qquad \text{for all } x \leq t \tag{10}$$

When (8) holds, (9) and (10) give

$$\int_{-\infty}^{\infty} u(x)dF(x) = \mu(F) - k\rho(F)$$

so that indifference curves in the space of feasible (μ, ρ) pairs will be parallel straight-line segments with positive slope, provided that φ is not uniformly zero. When the α-t mean-risk utility model is congruent with the expected utility model, (9) and (10) give

$$u(x) = x \qquad \text{for all } x \geq t \tag{11}$$

$$u(x) = x - k(t-x)^{\alpha} \qquad \text{for all } x \leq t \tag{12}$$

By (12), k is the unique solution to $u(t-1) = t - 1 - k$. In general, regardless of how an origin and scale unit are set for u, k is given by

$$k + 1 = \frac{u(t) - u(t-1)}{u(t+1) - u(t)}$$

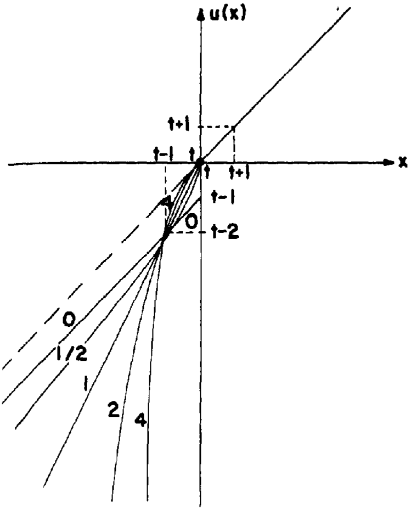


FIGURE 1 PLOTS OF $u(x)$ BY (11) AND (12) FOR $k = 1$ AND FIVE VALUES OF α

since this agrees with (11) and (12) and the ratio of u differences is invariant under a positive linear transformation of u. For example, $k = 1$ if and only if the difference in utility between t and $t - 1$ is twice the difference in utility between $t + 1$ and t.

Pictures of u according to (11) and (12) are shown in Figure 1 for $k = 1$ and for α values of 0, ½, 1, 2, and 4. These curves illustrate more clearly the types of behavior discussed after Theorem 1 although that discussion was not predicated on congruence with expected utility. All the functions are linear or risk neutral above t. The $\alpha = 0$ and $\alpha = 1$ curves are also linear below t but show a type of risk aversion around the target. The $\alpha = ½$ curve is convex or risk seeking below t, and the $\alpha = 2$ and $\alpha = 4$ curves are concave or risk averse below t. For more general discussions of risk attitudes in the expected utility setting, the reader can consult John Pratt and Kenneth Arrow.

An examination of von Neumann-Morgenstern utility functions which have been published in the investment literature gives mixed support for congruence between our version of mean-risk utility model and the expected utility model. Four individuals in Jackson Grayson's study on oil drilling decisions agree fairly well with (12) (see Albert Halter and Gerald Dean, p. 204) for net profit below a certain point, but exhibit a mixture of risk-seeking, risk-averse, and risk-neutral behavior above that point. The change point t below which (12) fits well is 0 for two individuals and approximately $-\$80,000$ and $-\$150,000$ for the other two. The drops below t are quite steep for each of the four (indicating relatively large k), and one has $\alpha \approx 1$ with $\alpha > 1$ for the other three.

Paul Green presents utility curves as functions of percent return on investment for four middle managers in a large chemical company. Equations (11) and (12) fit very well to three of the four curves and come close for the fourth. In each case t is 20 percent return on investment (an acknowledged goal for the firm) and utility is approximately linear above t. Risk-averse forms appear for returns below 20 percent, like for $\alpha = 2$ and $\alpha = 4$ on Figure 1. My estimates of the (α, k) pairs for (12) for the three best-fitting cases are (2.3, .16), (4.6, .00024), and (2.2, .19), so that α is larger in each case than the semivariance α of 2.

Ralph Swalm gives estimates of utility curves over large amounts of money for thirteen individuals engaged in a variety of enterprises. In so far as a change point t can be distinguished for these curves, it appears at or very near to zero, which is the no gain-no loss point in Swalm's setting. The linear form of (11) fits one curve very well, but ten of the thirteen are risk averse to varying degrees above zero, and the other two are slightly risk seeking. The predominant pattern below $t = 0$ is a slight amount of convexity, so that $\alpha < 1$ for most of the curves. The values of k appear to be smaller than those that apply to the oil and gas drilling decision makers in Grayson's study.

Halter and Dean (p. 64) sketch utility curves for changes in net wealth from $-\$50,000$ to $\$100,000$ for an orchard farmer, a grain farmer,

and a college professor. The orchard farmer's curve is linear above zero and below zero, but has a slope change at zero like $\alpha = 1$ in Figure 1. The grain farmer's curve has several inflection points above zero (from concave to convex back to concave) but is risk seeking below zero, indicating $\alpha < 1$. The college professor is risk averse above zero and risk seeking below zero with α slightly less than 1.

The general impression obtained from these studies is that most individuals in investment contexts do indeed exhibit a target return—which can be above, at, or below the point of no gain and no loss—at which there is a pronounced change in the shape of their utility functions, and that (12) can give a reasonably good fit to most of these curves in the below-target region. However, the linearity of (11) holds only in a limited number of cases for returns above target. A possibly optimistic estimate of the number of curves in the four studies cited above for which utility is approximately linear above the targets is 9 out of 24. The other 15 exhibit a variety of concave and convex forms above the targets with a tendency towards concavity or risk aversion. For below target utility, a wide range of α values were observed.

The preceding section showed how α for the α-t model could be estimated on the basis of (5) without presuming congruence between that model and the expected utility model. When congruence is assumed to hold, other methods for estimating α can be considered. For example, with $d > 0$, let d_1 be such that t as a sure-thing is indifferent to the 50–50 gamble for $t + d$ or $t - d_1$, and let d_2 be such that t is indifferent to the 50–50 gamble for $t + 2d$ or $t - d_2$. Then $2u(t) = u(t + d) + u(t - d_1) = u(t + 2d) + u(t - d_2)$, which on using (11) and (12) yields $\alpha = log[(2d - d_2)/(d - d_1)]/log(d_2/d_1)$. Needless to say, nonlinearities above t can affect this estimate, and it may be advisable to compare α values obtained from several values of d.

III. Stochastic Dominance Comparisons

Stochastic dominance criteria as discussed for example by Quirk and Saposnik, the author (1964), David Hertz, Josef Hadar and William Russell (1969, 1971), G. Hanoch and Levy, G. A. Whitmore, and Porter and Roger Bey, have attracted interest in investment decision making because of their correspondence with several types of von Neumann-Morgenstern utility functions and their ability to avoid certain criticisms that apply to the mean-variance dominance model. As far as I am aware, Porter was the first to demonstrate a close relationship between stochastic dominance and the mean-target semivariance dominance model. His result says that, except for cases of identical means and semivariances, if F dominates G by second degree stochastic dominance then F dominates G by the mean-target semivariance model. Hence, apart from the noted exception, the α-t efficient set for $\alpha = 2$ is a subset of the second degree stochastic dominance efficient set.

This section extends Porter's result to the general class of α-t models using first, second, and third degree stochastic dominance relations defined as follows:

F FSD G if and only if
$F \neq G$ and $F(x) \leq G(x)$ for all x

F SSD G if and only if
$F \neq G$ and $F_1(x) \leq G_1(x)$ for all x

F TSD G if and only if
$F \neq G$, $\mu(F) \geq \mu(G)$, and
$F_2(x) \leq G_2(x)$ for all x

where $F_\alpha(x)$ is the integral over y from $-\infty$ to x of $(x - y)^\alpha dF(y)$. Using integration by parts, we note that $F_1(x) = \int_{-\infty}^{x} F(y)\,dy$, so that $F_1(x)$ is the area under F up to x, and that $F_2(x) = 2\int_{-\infty}^{x} F_1(y)\,dy$, so that $F_2(x)$ is (twice) the area under F_1 up to x. It is easily verified that $F\,FSD\,G$ implies $F\,SSD\,G$ and that $F\,SSD\,G$ implies $F\,TSD\,G$ so that the TSD efficient set is a subset of the SSD efficient set which in turn is a subset of the FSD efficient set.

The connections between stochastic dominance and utility functions that will be used in looking at the α-t model are presented in the following lemma. For convenience we shall write

$E(u,F) = \int_{-\infty}^{\infty} u(x)\,dF(x)$. A real valued function v on returns is said to be concave (in the nonstrict sense) if $\lambda v(x) + (1-\lambda)v(y) \le v(\lambda x + (1-\lambda)y)$ for all x and y and all λ between 0 and 1. Proofs of the lemma are given in the references cited at the outset of this section.

LEMMA 1: *If F FSD G then $\mu(F) > \mu(G)$ and $E(u,F) \ge E(u,G)$ for every nondecreasing real valued function u;*

If F SSD G then $\mu(F) \ge \mu(G)$ and $E(u,F) \ge E(u,G)$ for every nondecreasing and concave real valued function u;

If F TSD G then $\mu(F) \ge \mu(G)$ and $E(u,F) \ge E(u,G)$ for every nondecreasing and concave real valued function u for which $-du/dx$ is concave.

When the appropriate derivatives exist, the functions used in the three parts of the lemma correspond respectively to $u' \ge 0$, $(u' \ge 0, u'' \le 0)$ and $(u' \ge 0, u'' \le 0, u''' \ge 0)$. Thus, for example, *SSD* corresponds to risk-averse utility functions.

THEOREM 3: *If F FSD G then $F\,P(\alpha,t)\,G$ for all $\alpha \ge 0$;*

If F SSD G then $F\,P(\alpha,t)\,G$ for all $\alpha \ge 1$, except when $\mu(F) = \mu(G)$ and $F_\alpha(t) = G_\alpha(t)$;

If F TSD G then $F\,P(\alpha,t)\,G$ for all $\alpha \ge 2$, except when $\mu(F) = \mu(G)$ and $F_\alpha(t) = G_\alpha(t)$.

This shows that the α-t efficient set is a subset of the *FSD* efficient set for every possible value of α; that the α-t efficient set is a subset of the *SSD* efficient set whenever $\alpha \ge 1$, except in the noted case; and that the α-t efficient set is a subset of the *TSD* efficient set whenever $\alpha \ge 2$, except in the noted case. The final observation shows that *SSD* can be replaced by *TSD* in Porter's semivariance result. A proof of Theorem 3 concludes the Appendix.

IV. Summary

This paper has examined a class of mean-risk dominance models in which risk equals the expected value of a function that is zero at and above a target return t and is nondecreasing in deviations below t. A special case of the general model is the α-t model in which distribution F dominates distribution G if $\mu(F) \ge \mu(G)$ and $\int_{-\infty}^{t} (t-x)^\alpha\,dF(x) \le \int_{-\infty}^{t} (t-x)^\alpha\,dG(x)$ with at least one strict inequality.

The model was motivated by the observation that decision makers in investment contexts frequently associate risk with failure to attain a target return. Examination of published utility functions from studies which use the maximization of expected utility criterion lends support to the notion of a target return at which the utility function undergoes a noticeable change. Depending on context, the change point may be negative, zero, or positive.

The α-t model describes choice by maximum expected value when all returns are above target, and it exhibits a tendency to avoid distributions having below-target returns. Choice between a risky option, all of whose possible returns are below the target, and a sure-thing option yielding the mean of the risky option depends on the value of α. If $\alpha < 1$ then the risky option dominates the sure-thing, and if $\alpha > 1$ then the sure-thing dominates the risky option.

If the α-t utility model—which presumes the existence of a real valued function U in mean and risk which increases in mean, decreases in risk, and reflects the decision maker's preferences between distributions—is congruent with the von Neumann-Morgenstern expected utility model with utility function u, then u can be written as $u(x) = x$ for $x \ge t$, $u(x) = x - k(t-x)^\alpha$ for $x \le t$, with $k > 0$. Utilities for returns below target in published studies can be approximated rather well by $x - k(t-x)^\alpha$. Values of α ranging from less than 1 to greater than 4 were observed. A minority of utility functions examined agreed with the linearity of utility above t. Those that were not linear above t exhibited a variety of risk-averse and risk-seeking shapes.

Methods of estimating α, with and without presumption of congruence between the α-t model and the expected utility model, were presented. Finally, it was shown that, except for distributions with equal means and equal risks, the efficient set for the α-t model is a sub-

set of the *FSD* efficient set for all $\alpha \geq 0$, the α-t efficient set is a subset of the *SSD* efficient set for all $\alpha \geq 1$, and the α-t efficient set is a subset of the *TSD* efficient set for all $\alpha \geq 2$.

APPENDIX

The proof of Theorem 1(c) follows readily from a strict inequality version of Jensen's inequality which says that if G is a nondegenerate distribution (in the sense that G assigns positive probability to the event that the return is not $\mu(G)$) having probability 1 on a finite interval I, and if f is a strictly concave function on I, then $\int_I f(x)\,dG(x) < f(\mu(G))$. With F and G as specified in (c), let $f(x) = (t-x)^\alpha$ for $x \leq t$. If $0 < \alpha < 1$ then f is strictly concave, and if $\alpha > 1$ then $-f$ is strictly concave. Thus if $0 < \alpha < 1$ then $\int_{-\infty}^{t} (t-x)^\alpha\,dG(x) < d^\alpha$, and if $\alpha > 1$ then $d^\alpha < \int_{-\infty}^{t} (t-x)^\alpha\,dG(x)$ by Jensen's inequality. Since $d^\alpha = \int_{-\infty}^{t} (t-x)^\alpha\,dF(x)$, $0 < \alpha < 1$ gives $G_\alpha(t) < F_\alpha(t)$ which along with $\mu(F) = \mu(G)$ implies $G\,P(\alpha,t)F$; and $\alpha > 1$ implies $F\,P(\alpha,t)G$ in like manner. If $\alpha = 1$ then $G_\alpha(t) = F_\alpha(t)$ and neither F nor G stands in the $P(\alpha,t)$ relation to the other.

For notational simplicity in proving Theorem 2, set $t = 0$, with $u(0) = 0$ and $u(1) = 1$. For $x > 1$, the gamble with probability $1/x$ for x and $1 - 1/x$ for 0 has $\mu = 1$ and $\rho = 0$, so that its U value is $U(1,0)$. This is also the U value of the gamble that yields 1 with probability 1. Hence, by (8), $(1/x)u(x) + (1 - 1/x)u(0) = u(1)$, so that $u(x) = x$. For $0 < x < 1$, the gamble with probability $1/(2-x)$ for x and $(1-x)/(2-x)$ for 2 has $(\mu, \rho) = (1, 0)$, so that $[1/(2-x)]u(x) + [(1-x)/(2-x)]u(2) = u(1)$, again giving $u(x) = x$. Thus, (9) holds.

If φ is uniformly zero then (10) holds in similar fashion. Assume henceforth that $\varphi(y)$, as a nonnegative nondecreasing function in y with $\varphi(0) = 0$, is positive for some y. For definiteness we presume that $\varphi(1) > 0$ and let $k = -[1 + u(-1)]/\varphi(1)$. Comparing the 50-50 gamble for 1 or -1 with the sure-thing for 0, (8) and the specification following (8) that $U(0,0) > U(0,\varphi(1)/2)$ imply $0 > u(1)/2 + u(-1)/2$, or $0 > 1 + u(-1)$ so that $k > 0$. For $x \leq -1$ let w exceed $-x$, define $z = (x-w)\varphi(1)/\varphi(-x) + 2w + 1$, take F_1 as the 50-50 gamble for -1 or z, and let F_2 be the distribution which has probability $\varphi(1)/[2\varphi(-x)]$ for x and probability $[2\varphi(-x) - \varphi(1)]/[2\varphi(-x)]$ for w. Then $\mu(F_1) = \mu(F_2)$ and $\rho(F_1) = \rho(F_2)$ so that (8) gives

$$\frac{1}{2}u(-1) + \frac{1}{2}u(z) = \frac{\varphi(1)}{2\varphi(-x)}u(x) + \frac{2\varphi(-x) - \varphi(1)}{2\varphi(-x)}u(w)$$

Since $u(z) = z$ and $u(w) = w$ by (9), substitution in the preceding equation and the definition of z given above yield $u(x) = x - k\varphi(-x)$.

Finally, for $-1 < x < 0$ let w exceed $-x$, define $z = [\varphi(1)(x + w) + \varphi(-x)]/[2\varphi(1) - \varphi(-x)]$, take F_1 as the 50-50 gamble for x or w, and let F_2 be the distribution with probability $\varphi(-x)/[2\varphi(1)]$ for -1 and probability $[2\varphi(1) - \varphi(-x)]/[2\varphi(1)]$ for z. Then $\mu(F_1) = \mu(F_2)$ and $\rho(F_1) = \rho(F_2) = \varphi(-x)/2$ so that (8) gives

$$\frac{1}{2}u(x) + \frac{1}{2}u(w) = \frac{\rho(-x)}{2\rho(1)}u(-1) + \frac{2\rho(1) - \rho(-x)}{2\rho(1)}u(z)$$

Since $u(z) = z$ and $u(w) = w$ by (9), substitution in the preceding equation and the definition of z yield $u(x) = x - k\varphi(-x)$. This completes the proof of Theorem 2.

The proof of Theorem 3 is straightforward. Let $v(x) = -[\max\{0, t - x\}]^\alpha$ for all x. When $\alpha \geq 0$, v is nondecreasing in x; when $\alpha \geq 1$, v is nondecreasing and concave; and when $\alpha \geq 2$, v satisfies the conditions in the final part of the lemma. Hence $F_\alpha(t) \leq G_\alpha(t)$ whenever $F\,FSD\,G$ and $\alpha \geq 0$; $F_\alpha(t) \leq G_\alpha(t)$ whenever $F\,SSD\,G$ and $\alpha \geq 1$; and $F_\alpha(t) \leq G_\alpha(t)$ whenever $F\,TSD\,G$ and $\alpha \geq 2$. Theorem 3 then follows immediately from these observations and the inequalities involving μ in Lemma 1.

REFERENCES

C. P. Alderfer and H. Bierman, "Choices with Risk: Beyond the Mean and Variance," *J. Bus., Univ. Chicago,* July 1970, *43,* 341–53.

Kenneth J. Arrow, *Aspects of the Theory of Risk-Bearing,* Helsinki 1965.

W. J. Baumol, "An Expected Gain-Confidence Limit Criterion for Portfolio Selection," *Manage. Sci.,* Oct. 1963, *10,* 174–82.

G. O. Bierwag, "Liquidity Preference and Risk Aversion with an Exponential Utility Function: Comment," *Rev. Econ. Stud.,* Apr. 1974, *41,* 301–02.

——— **and M. A. Grove,** "Indifference Curves in Asset Analysis," *Econ. J.,* June 1966, *76,* 337–43.

K. Borch, "A Note on Utility and Attitudes to Risk," *Manage. Sci.,* July 1963, *9,* 697–700.

———, "Indifference Curves and Uncertainty," *Swedish J. Econ.,* Mar 1968, *70,* 19–24.

———, "A Note on Uncertainty and Indifference Curves," *Rev. Econ Stud.,* Jan. 1969, *36,* 1–4.

———, "The Rationale of the Mean-Standard Deviation Analysis: Comment," *Amer. Econ. Rev.,* June 1974, *64,* 428–30.

J. S. Chipman, "The Ordering of Portfolios in Terms of Mean and Variance," *Rev. Econ. Stud.,* Apr. 1973, *40,* 167–90.

E. V. Domar and R. A. Musgrave, "Proportional Income Taxation and Risk-Taking," *Quart J. Econ.,* May 1944, *58,* 389–422.

M. S. Feldstein, "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Rev. Econ. Stud.,* Jan. 1969, *36,* 5–12.

Peter C. Fishburn, *Decision and Value Theory,* New York 1964.

———, "On the Foundations of Mean-Variance Analyses," mimeo, 1975.

C. J. Grayson, *Decisions Under Uncertainty: Drilling Decisions by Oil and Gas Operators,* Graduate School of Business, Harvard Univ. 1960.

P. E. Green, "Risk Attitudes and Chemical Investment Decisions," *Chem. Eng. Progress,* Jan. 1963, *59,* 35–40.

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.,* Mar. 1969, *59,* 25–34.

——— **and** ———, "Stochastic Dominance and Diversification," *J. Econ. Theory,* Sept. 1971, *3,* 288–305.

Albert N. Halter and Gerald W. Dean, *Decisions Under Uncertainty,* Cincinnati 1971.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.,* July 1969, *36,* 335–46.

D. B. Hertz, "Investment Policies that Pay Off," *Harvard Bus. Rev.,* Jan./Feb. 1968, *46,* 96–108.

W. W. Hogan and J. M. Warren, "Computation of the Efficient Boundary in the E-S Portfolio Selection Model," *J. Finance. Quant. Anal.,* Sept. 1972, *7,* 1881–96.

——— **and** ———, "Toward the Development of an Equilibrium Capital-Market Model Based on Semivariance," *J. Finance. Quant. Anal.,* Jan. 1974, *9,* 1–11.

A. K. Klevorick, "A Note on 'The Ordering of Portfolios in Terms of Mean and Variance,'" *Rev. Econ. Stud.,* Apr. 1973, *40,* 293–96.

H. Levy, "The Rationale of the Mean-Standard Deviation Analysis: Comment," *Amer. Econ. Rev.,* June 1974, *64,* 434–41

J. Lintner, (1965a) "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.,* Feb. 1965, *47,* 13–37.

———, (1965b) "Security Prices, Risk, and Maximal Gains from Diversification," *J. Finance,* Dec. 1965, *20,* 587–615.

J. C. T. Mao, (1970a) "Survey of Capital Budgeting: Theory and Practice," *J. Finance,* May 1970, *25,* 349–60.

———, (1970b) "Models of Capital Budgeting, E-V vs. E-S," *J. Finance. Quant. Anal.,* Jan. 1970, *4,* 657–75.

Harry Markowitz, "Portfolio Selection," *J. Finance,* Mar. 1952, *7,* 77–91.

———, *Portfolio Selection,* New York 1959.

R. B. Porter, "Semivariance and Stochastic Dominance: A Comparison," *Amer. Econ. Rev.*, Mar. 1974, *64*, 200–04.

——— **and R. P. Bey,** "An Evaluation of the Empirical Significance of Optimal Seeking Algorithms in Portfolio Selection," *J. Finance*, Dec. 1974, *29*, 1479–90.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

D. G. Pruitt, "Pattern and Level of Risk in Gambling Decisions," *Psychological Rev.*, May 1962, *69*, 187–201.

J. P. Quirk and R. Saposnik, "Admissibility and Measurable Utility Functions," *Rev. Econ. Stud.*, Feb. 1962, *29*, 140–46.

P. A. Samuelson, "General Proof that Diversification Pays," *J. Finance. Quant. Anal.*, Mar. 1967, *2*, 1–13.

———, "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments," *Rev. Econ. Stud.*, Oct. 1970, *37*, 537–42.

——— **and R. C. Merton,** "Generalized Mean-Variance Tradeoffs for Best Perturbation Corrections to Approximate Portfolio Decisions," *J. Finance*, Mar. 1974, *29*, 27–40.

W. F. Sharpe, "A Simplified Model for Portfolio Analysis," *Manage. Sci.*, Jan. 1963, *9*, 277–93.

———, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.

B. K. Stone, "A General Class of Three-Parameter Risk Measures," *J. Finance*, June 1973, *28*, 675–85.

R. D. Swalm, "Utility Theory—Insights into Risk Taking," *Harvard Bus. Rev.*, Nov./Dec. 1966, *47*, 123–36.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–85.

———, "The Theory of Portfolio Selection," in F. H. Hahn and F. P. R. Brechling, eds., *The Theory of Interest Rates*, London 1965, 3–51.

S. C. Tsiang, "The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money," *Amer. Econ. Rev.*, June 1972, *62*, 354–71.

———, "The Rationale of the Mean-Standard Deviation Analysis: Reply and Errata for Original Article," *Amer. Econ. Rev.*, July 1974, *64*, 442–50.

John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, 2d ed., Princeton 1947.

G. A. Whitmore, "Third-Degree Stochastic Dominance," *Amer. Econ. Rev.*, June 1970, *60*, 457–59.

Quality Choice and Competition

By HAYNE E. LELAND*

Current economic theory does not offer a comprehensive explanation of quality choices by firms. For *fixed* product designs or quality levels, received theory explains quantity decisions made both by perfectly and by imperfectly competitive firms. But production designs typically are not fixed. Firms must choose not only how much they produce, but also the nature of *what* they produce. By altering design parameters and inputs per unit output, firms can choose from a wide range of alternative "quality" levels: any parking lot provides sufficient evidence that a car is not a car is not a car. And in their effect upon consumer welfare, the decisions about what is produced are as important as the decisions about how much is produced.

Traditional theory has treated different quality levels of a good as if they were different goods. The problem with this approach is that there is no metric to determine the "closeness" of different products. Without such a measuring rod, it is difficult to use the powerful tools of continuity and of marginal analysis, which have proved so useful in developing the theory of quantity selection by firms.

To bypass this problem, Robert Dorfman and Peter Steiner, James Rosse, and others have introduced quality parameters directly into demand functions: $D = D(p, q)$, where p is price and q is quality. This permits marginal analysis. But it can address welfare questions only when combined with the strong assumptions justifying a "consumer surplus" framework. Hendrik Houthakker examined consumer theory with a single quality variable introduced directly into the utility function. He did not consider properties of equilibrium, however.

In this paper, I develop a comprehensive theory of quality choice using an alternative approach to consumer theory. This approach, developed by Kelvin Lancaster (1966), presumes that goods themselves do not directly provide utility, but rather provide basic "characteristics" which consumers value. Related approaches have been used by Franklin Fisher, Zvi Griliches, and Carl Kaysen to study automobile quality change, and by Richard Muth to examine housing demand. The "household production function" approach of Gary Becker and others utilizes similar concepts. These papers have emphasized consumer behavior but have not examined properties of market equilibrium. Sherwin Rosen recently developed a partial equilibrium model of quality choice. His approach is different from mine in that it examines a single market in which a continuum of different quality levels exist. Rosen addresses a rather different set of questions than that considered in this paper.

The characteristics approach describes each good by an S dimensional vector, whose elements indicate the amount of each characteristic provided per unit of that good. One advantage of this analysis is that it introduces a natural metric on the closeness of goods. Another advantage is that it permits a simple notion of quality change: a change in the amount of each

*Associate professor, Graduate School of Business Administration, University of California, Berkeley. Research for this paper was supported in part by a grant from the Dean Witter Foundation. An earlier version of this paper appeared as working paper no. 48, IMSSS, Stanford University, September 1974. I wish to thank members of the Berkeley Applied Economics Workshop for useful suggestions. Since this paper was written, two related papers have come to my attention. Jacques Drèze and Kåre Hagen's paper takes a similar approach, although the authors emphasize problems of nonconvexity and do not focus on the key role of implicit characteristic prices; and Michael Spence uses a consumer surplus approach to examine monopolistic competition and the number and quantity of goods of different quality produced. At the cost of losing some generality of consumer preferences, Spence's framework can examine the difficult problem of whether the "optimal" number of firms and therefore quality levels exist.

characteristic provided per unit of the good.[1] Note that "quality change" by this definition may not be unambiguously good or bad. Some characteristics may be provided in greater quantities, while others in lesser. Individuals may differ in whether they regard such a change as desirable. We thus avoid some of the conceptual problems encountered by James Sweeney and others in defining quality.

While the characteristics formulation is familiar to students of consumer theory, many of my results draw from analysis in quite a different field: asset market equilibrium and production under uncertainty. The similarities are based on the fact that a share of stock can be viewed as a vector of returns across states of nature (rather than across characteristics). As firms change production decisions, this pattern of returns will change (i.e., there is a change in quality of returns). Results on the optimality of production under uncertainty have been developed by the author. Some of what follows involves a reinterpretation of these results in the framework of characteristics and quality choice.

In Sections I–III, a simple model of general equilibrium with quality choice is developed. Section IV generates necessary conditions for quality and quantity choices to be Pareto optimal. Sections V and VI introduce two concepts which are crucial to optimality: "spanning" and "competitive implicit characteristic prices." Section VII proves that these properties are sufficient for profit-maximizing firms to make decisions which satisfy the Pareto optimality conditions developed in Section IV, and shows that quality choice has aspects of a "public good." Section VIII considers Pareto optimality when spanning is not present. Section IX examines divergences from competitive behavior and indicates that monopolies tend to underprovide quality, given their output choice.

I. Goods and Firms

Let us assume that there is a finite number of characteristics $s = 1, \ldots, S$ which generate utility for at least one consumer. A good j is described by a vector

$$c^j = (c_1^j, \ldots, c_S^j) \qquad j = 1, \ldots, J$$

where c_s^j is the amount of characteristic s provided per unit of good j. For simplicity, it is assumed each good can be associated with a corresponding firm $j = 1, \ldots, J$.[2]

Quality variations can now be parameterized simply. Let q^j be a design or quality parameter which can be adjusted by the firm.[3] Changing q^j will in general change each element of the vector c^j Thus a good with variable quality is described by[4]

$$c^j(q^j) = [c_1^j(q^j), \ldots, c_S^j(q^j)]$$

It is not required that individuals react unanimously to changes in the quality variable. If an increase in q^j increases the provision of all characteristics (which are presumed to be desirable) then consumers will agree unanimously that "quality has increased." But if some characteristics are enhanced whereas others are diminished, consumers may disagree as to whether quality has increased or decreased.

Firms in the model are characterized by an implicit production function

$$(1) \qquad f^j(q^j, y^j, x^j) = 0$$

which relates quality q^j and output quantity y^j with input quantities $x^j = (x_1^j, \ldots, x_L^j)$. We shall assume that f^j exhibits the usual concavity

[1]The reader may question the distinction between different goods versus different quality levels of the same good Both have the property that their vectors of characteristics provided are different. To make the distinction useful, we say two goods are different (rather than being different quality levels of the same good) if it is impossible for a firm which is set up to produce one good to change its design or quality parameter to produce the other. The distinction is not vital, however, for the analysis which follows

[2]This assumption is not restrictive given a fixed number of firms. But it does preclude a full analysis of entry and the question of whether the market provides a sufficient *diversity* of quality. See Section IX for further discussion

[3]For convenience, we shall assume $q^j \in Q^j$, where $Q^j \subset R^1$ More general approaches, such as $Q^j \subset R^N$, can easily be developed

[4]Note that externalities are ruled out by assuming only quality parameter q^j affects the provision of characteristics by good j.

properties with respect to q^j as well as with respect to y^j and x^j. Note that for a fixed input vector x^j, (1) describes a transformation curve between maximal output y^j and quality q^j per unit output.

Firms will choose the (q^j, y^j, x^j) combination satisfying (1) which maximizes profits. We presume initially that output price p^j may depend upon y^j and q^j. For simplicity, we assume the vector $r = (r_1, \ldots, r_L)$ of input prices is considered constant. Thus

$$(2) \quad \pi^j(q^j, y^j, x^j) = p^j(q^j, y^j)y^j - r'x^j$$

Maximizing (2) subject to (1) with respect to (q^j, x^j, y^j) yields first-order necessary conditions

$$(3) \quad \frac{\partial p^j(q^j, y^j)}{\partial q^j} y^j + \mu^j f^j_q = 0$$

$$(4) \quad p^j(q^j, y^j) + \frac{\partial p^j(q^j, y^j)}{\partial y^j} y^j + \mu^j f^j_y = 0$$

$$(5) \quad -r + \mu^j f^j_x = 0$$

$$(6) \quad f^j(q^j, y^j, x^j) = 0 \qquad j = 1, \ldots, J$$

where $f^j_q = \partial f^j(q^j, y^j)/\partial q^j$, etc., and μ^j is the Lagrangean multiplier associated with the constraint. Note that if q^j is fixed, equations (4)–(6) describe the usual profit-maximizing conditions.

II. Consumers and Demand

Consumers are assumed to have preferences defined over bundles of characteristics. Let $R_i = (R_{i1}, \ldots, R_{iS})$ describe a bundle consumed by i, where R_{iS} is the amount of characteristic s. If preference rankings over a set of bundles exhibit the normal properties, consumers will possess *nonsatiated* quasi-concave ordinal utility functions

$$(7) \quad U_i = U_i(R_{i1}, \ldots, R_{iS}) \qquad i = 1, \ldots, I$$

This formulation is consistent both with Lancaster's approach (where s indexes characteristics) and with the "state preference" approach of Arrow-Debreu (where s indexes states of nature and R_i is a vector of returns across states).

The amount of each characteristic s consumed by i depends on the bundle of goods consumed and the quality of those goods. Following Lancaster's emphasis that characteristics are in principle physically measurable, we presume that the amount of characteristic s a bundle of goods $y_i = (y_i^1, \ldots, y_i^J)$ provides is equal to the sum of the contributions of each unit of each good.[5] Thus

$$(8) \quad R_{is} = \sum_j c^j_s(q^j)y^j_i \qquad s = 1, \ldots, S$$

or in matrix terms

$$(8') \qquad R_i = C(q)y_i$$

where $C(q)$ is an $S \times J$ matrix with elements $c^j_s(q^j)$ and q is the vector of quality decisions $q^1, \ldots, q^J$.[6]

Consumers select the bundle of goods y_i which maximizes utility (7) subject to a budget constraint

$$(9) \qquad \sum_j p^j y^j_i = \sum_\ell r_\ell \bar{x}_{i\ell} + \sum_j \bar{\theta}^j_i \pi^j$$

or in vector notation

$$(9') \qquad p'y_i = r'\bar{x}_i + \bar{\theta}'_i \pi$$

where $p = (p^1, \ldots, p^J)$ is the vector of prices of goods $1, \ldots, J$; $\bar{x}_i = (\bar{x}^1_i, \ldots, \bar{x}^L_i)$ is the vector of primary goods initially owned by consumer i; and $\bar{\theta}_i = (\bar{\theta}^1_i, \ldots, \bar{\theta}^J_i)$ is the vector of fractions of each firm owned by consumer i. Transposes denote row vectors. Appending the budget constraint with Lagrangean multiplier λ_i and finding a stationary point gives first-order necessary conditions for utility maximization:

$$(10) \quad \sum_s U_{is} c^j_s(q^j) - \lambda_i p^j = 0 \quad j = 1, \ldots, J; \quad i = 1, \ldots, I$$

where $U_{is} = \partial U_i / \partial R_{is}$, or in matrix terms

$$(10') \qquad U'_i C(q) - \lambda_i p' = 0 \qquad i = 1, \ldots, I$$

[5]More generally, we could allow for possible interactions through the introduction of consumption "activities," as in Lancaster (1966). For simplicity, we use the simpler approach embodied in (8).

[6]Note that factors $x = (x_1, \ldots, x_L)$ are assumed not to affect utilities. This assumption could easily be relaxed.

where[7]

$$U_i' = (U_{i1}, \ldots, U_{iS})$$

It is of some interest to note the similarity between conditions (10) and the portfolio equilibrium conditions in my earlier paper. In fact, the conditions are formally identical. It is often useful to think of a commodity bundle as a "portfolio," providing an optimal balance of "returns" across characteristics. Further axioms on choice could lead to an equivalent of the expected utility theorem, with the corpus of portfolio theory becoming directly applicable. Such is not, however, my current purpose. The immediate goal is to characterize quality choices by firms in general equilibrium.

III. Equilibrium

Equilibrium requires that individual units have no motivation to change their decisions and that markets are cleared. In the model, consumers will be in equilibrium when conditions (9) and (10) are satisfied. Note that consumers regard quality and price vectors as parameters.

Firms will be in equilibrium given conditions (3)–(6) are satisfied. Note that the derivatives $\partial p^j/\partial q^j$ and $\partial p^j/\partial y^j$ are *perceived* price changes, which may or may not be those which actually would occur. Models of perfect competition, for example, assume that $\partial p^j/\partial y^j = 0$. Since quality changes are not considered, this is equivalent to assuming that firms view prices as parameters.

But surely—even if $\partial p^j/\partial y^j = 0$—it cannot be argued that $\partial p^j/\partial q^j = 0$. In fact, if there were no perceived relation between price and quality, firms would be motivated to produce only the least expensive (and perhaps lowest quality) good. So a critical question affecting equilibrium and its optimality (or lack thereof) is specifying how firms perceive $\partial p^j/\partial q^j$. This is explored in detail in Section VII.

The final link to close the equilibrium model is a set of market equilibrium conditions

$$(11) \quad \sum_i y_i^j = y^j \qquad j = 1, \ldots, J$$

$$(12) \quad \sum_j x_\ell^j = \sum_i \bar{x}_{i\ell} \qquad \ell = 1, \ldots, L$$

Equilibrium, then, must satisfy equations (3)–(6) and (9)–(12). Since our present concern is with properties of equilibrium when it exists, I shall not address the difficult problem of proving that an equilibrium does exist.[8]

IV. Pareto Optimality with Quality Choice: Necessary Conditions

To examine necessary conditions for Pareto optimality, I use the standard technique of maximizing the utility of an arbitrary "first" consumer, holding other utility levels constant. For ease of notation, let us consider the two-consumer case with a single input ($I = 2$, $L = 1$). The diligent reader can ascertain that the results hold in the general case. It should be noted that the *number* of firms is treated as an exogenous variable.[9]

Optimality requires that quality, production, and distribution decisions

Maximize $U_1(R_{11}, \ldots, R_{1S})$
subject to

(a) $U_2(R_{21}, \ldots, R_{2S}) = \bar{U}_2$

(b) $y_1^j + y_2^j = y^j \qquad j = 1, \ldots, J$

(c) $\sum_j x^j = \bar{x} = \bar{x}_1 + \bar{x}_2$

(d) $f^j(q^j, y^j, x^j) = 0 \qquad j = 1, \ldots, J$

The maximization is with respect to y_1^j, y_2^j, y^j,

[7]We make the strong assumption that all choices are made in the interior of choice sets. That is, we assume $y_i^j > 0$ for all i and j. The Appendix discusses modifications required when corner solutions exist

[8]Under assumptions of spanning and competitive characteristic prices, there is similarity between this model and that considered by Gerard Debreu. Roy Radner proves this equivalence in the context of capital asset market equilibrium with production

[9]Thus the results shed no light on the optimality of the number of different goods (and different quality levels) provided by the system. Lancaster (1975) has made initial progress in analyzing this question in a simplified framework, as has Michael Spence. Our focus is on whether quality decisions by firms are optimal, *given* the number of firms (and therefore the number of different goods) is fixed

x^j, and q^j. Rather than append constraint (b), we substitute for $y_2^j = y^j - y_1^j$ directly. The Lagrangean expression is

$$L = U_1\Big[\sum_j c_1^j(q^j)y_1^j, \ldots, \sum_j c_S^j(q^j)y_1^j\Big] + \lambda^*\Big\{U_2\Big[\sum_j c_1^j(q^j)(y^j - y_1^j), \ldots, \sum_j c_S^j(q^j)(y^j - y_1^j)\Big] - \bar{U}_2\Big\} + \gamma^*\Big[\sum_j x^j - \bar{x}\Big] + \sum_j \mu^{j*} f^j(q^j, y^j, x^j)$$

Stationary conditions are

$$(13)\quad \frac{\partial L}{\partial y_1^j} = \sum_s U_{1s} c_s^j(q^j) - \lambda^* \sum_s U_{2s} c_s^j(q^j) = 0$$

$$(14)\quad \frac{\partial L}{\partial y^j} = \lambda^* \sum_s U_{2s} c_s^j(q^j) + \mu^{j*} f_y^j = 0$$

$$(15)\quad \frac{\partial L}{\partial x^j} = \gamma^* + \mu^{j*} f_x^j = 0$$

$$(16)\quad \frac{\partial L}{\partial q^j} = \sum_s U_{1s} \frac{\partial c_s^j}{\partial q^j} y_1^j + \lambda^* \sum_s U_{2s} \frac{\partial c_s^j}{\partial q^j} (y^j - y_1^j) + \mu^{j*} f_q^j = 0$$

$$(17)\quad \frac{\partial L}{\partial \mu^{j*}} = f^j(q^j, y^j, x^j) = 0$$

$$(18)\quad y_1^j + y_2^j = y^j$$

$$(19)\quad \frac{\partial L}{\partial \gamma^*} = \sum_j x^j - \bar{x} = 0$$

where equations (13)–(18) hold for $j = 1, \ldots, J$. We also have $U_2 = \bar{U}_2$, but since $\bar{U}_2$ can be set arbitrarily, it is not included directly with the other necessary conditions.

These conditions will be sufficient as well as necessary (assuming an equilibrium exists) if utility functions are jointly quasi concave in y_i^j and q^j. Unfortunately, this requirement need not always be satisfied. Utility functions involve arguments of the form $c_s^j(q^j)y_i^j$. Even if $c_s^j(q^j)$ is strictly concave, the product of the terms will not necessarily be jointly quasi concave in q^j and y_i^j.[10]

In general, the optimality conditions (13)–(19) will not be satisfied by the equilibrium developed in previous sections. This is hardly surprising, since received theory indicates prices must be regarded as invariant to output ($\partial p^j/\partial y^j = 0$) for Pareto optimality.

Even if we restricted attention to the case where $\partial p^j/\partial y^j = 0$, there remains the term $\partial p^j/\partial q^j$—the jth firm's perception of how its price responds to quality change. We are reduced to the essential question: what further restrictions on competitive behavior must be satisfied if the equilibrium is to satisfy the necessary conditions for Pareto optimality?

My earlier paper examined conditions under which firms would choose Pareto optimal patterns of returns across states of nature. Two properties were chosen to be crucial to the optimality of asset market equilibrium: "spanning," and "competitive implicit contingency claim" prices. Both concepts are equally important to the optimality of quality choices by firms, and are examined in the following sections.

V. Spanning

Changes in patterns of characteristics consumed can occur in two ways. First, the consumer can alter the portfolio of goods he consumes. Such a portfolio change has a well-defined cost (perhaps negative). Second, firms can alter the quality of the goods they produce, thereby changing the pattern of returns to fixed bundles of goods. The personal value of such a quality change, per unit of the good consumed, will in general differ among consumers. But if every change in pattern of returns resulting from quality change can be duplicated by a change in portfolio, there will exist a common money "value" for quality change—namely, the value of the corresponding portfolio change.

Essentially, the spanning property says that any small change in characteristics effected by

[10]A similar point has been made by Jacques Drèze in the context of uncertainty

quality change can be effected by some portfolio change of the goods consumed.[11] Mathematically, spanning implies the existence of vectors

$$h^j(q) = [h_1^j(q), \ldots, h_J^j(q)] \qquad j = 1, \ldots, J$$

such that

$$(20) \qquad \frac{\partial c_s^j(q^j)}{\partial q^j} = \sum_{k=1}^{J} c_s^k(q^k) h_k^j(q)$$

$$s = 1, \ldots, S \qquad j = 1, \ldots, J$$

or in matrix terms,

$$(20') \qquad c_q^j(q^j) = C(q) h^j(q)$$

where $c_q^j(q^j)$ is an S dimensional vector with elements $\partial c_s^j(q^j)/\partial q^j$.

Consider now the change in price $\partial p_i^j/\partial q^j$ that consumer i would just be willing to pay for a small change in the quality parameter of firm j. Clearly, $\partial p_i^j/\partial q^j$ will be the price change which renders $\partial U_i/\partial q^j = 0$. Equally clearly, $\partial p_i^j/\partial q^j$ will in general differ between consumers. But we show below that *spanning implies that $\partial p_i^j/\partial q_j$ is the same for all consumers*.

Appending the budget constraint (9) to (7) with the Lagrangean multiplier λ_i satisfying (10), and differentiating the resulting expression with respect to q^j gives

$$(21) \quad \frac{\partial U_i}{\partial q^j} = \sum_s U_{is}\left(\frac{\partial c_s^j}{\partial q^j} y_i^j + \sum_k c_s^k(q^k)\frac{\partial y_i^k}{\partial q^j}\right) - \lambda_i\left(\sum_k p^k \frac{\partial y_i^k}{\partial q^j} + \frac{\partial p_i^j}{\partial q^j} y_i^j\right) = 0$$

or using (10),

$$\left[\sum_s U_{is}(\partial c_s^j/\partial q^j) - \lambda_i(\partial p_i^j/\partial q^j)\right] y_i^j = 0$$

From the spanning condition (20), we may rewrite (21) as

$$(22) \quad \left[\sum_s U_{is} \sum_k c_s^k(q^k) h_k^j(q) - \lambda_i \frac{\partial p_i^j}{\partial q^j}\right] y_i^j = 0$$

$$\left[\sum_k \sum_s U_{is} c_s^k(q^k) h_k^j(q) - \lambda_i \frac{\partial p_i^j}{\partial q^j}\right] y_i^j = 0$$

$$\left[\sum_k p^k h_k^j(q) - \frac{\partial p_i^j}{\partial q^j}\right] \lambda_i y_i^j = 0$$

again using (10). Our assumptions of nonsatiation and interior solutions therefore imply that

$$(23) \qquad \frac{\partial p_i^j}{\partial q^j} = \sum_k p^k h_k^j(q)$$

The remarkable aspect of (23) is that the right side is independent of i Therefore *spanning ensures that the price change that each consumer would be willing to pay for a small change in quality is identical for all consumers.*[12] If spanning is not satisfied, there will exist consumers who value quality changes at different prices per unit.[13]

One can see why spanning may be a necessary condition for Pareto optimality. If rates of substitution between quality and income (and therefore between quality and inputs x) differ, trading with markets may not lead to optimal decisions. Some consumers will want more quality, others less. Further bargaining between individuals, as contrasted with trading in markets, may be required for optimality. We examine this question rigorously in Sections VIII and IX.

Is spanning likely to be satisfied? There are a number of situations which imply spanning in capital asset markets (see my earlier paper, Propositions I–IV). One of these seems particularly relevant to this study: the case of "*complete markets*," when the number of goods with linearly independent vectors of characteristics is as great as the number of characteristics ($J \geqq S$).

[11]Of course, the cost of the change in characteristics resulting from a quality change may differ from the cost of the spanning portfolio of goods. In a Pareto optimal equilibrium, it can be shown that the two costs will be equal

[12]See the Appendix for how this conclusion may be modified when first-order conditions are satisfied by strict inequalities; i.e., corner solutions Our equation (21) also assumes that consumers perceive no change in their incomes when q^j changes, as will be the case when firms are in equilibrium ($\partial\pi^j/\partial q^j = 0$)

[13]If consumers have arbitrary vectors $U_i' = (U_{i1}, \ldots, U_{iS})$. If there are restrictions on tastes, spanning will not be a necessary condition for unanimity, although it clearly will remain sufficient

With complete markets, the spanning property will be satisfied. Since $C(q)$ will be of rank S, it will possess a right inverse—that is, there exists a matrix $A(q)$ such that

$$C(q)A(q) = I_S \tag{24}$$

where I_S is the identity matrix with rank S. Therefore, the vector $h^j(q) = A(q)c_q^j(q^j)$ exists, and since

$$C(q)h^j(q) = C(q)A(q)c_q^j(q^j) = c_q^j(q^j)$$

we have from (20′) that spanning is satisfied. Of course, spanning of the c_q^j vectors may still be satisfied if there are fewer goods than characteristics; for example, when a change in quality increases the vector of characteristics proportionately (which would be equivalent to a quantity increase), c_q^j is spanned by $C(q)$. But spanning will be satisfied for *arbitrary* c_q^j with complete markets.

VI. Competitive Implicit Characteristic Prices

From conditions (10), we have for all consumers $i = 1, \ldots, I$:

$$\sum_s \left(\frac{U_{is}}{\lambda_i}\right) c_s^j(q^j) = p^j \qquad j = 1, \ldots, J \tag{25}$$

or

$$\sum_s v_{is} c_s^j(q^j) = p^j \qquad j = 1, \ldots, J \tag{26}$$

where $v_{is} \equiv U_{is}/\lambda_i$.

Just as v_{is} could be interpreted as an implicit contingency claim price in the context of capital asset market equilibrium, so also can it be interpreted in the context of equilibrium with quality choice In the present case, v_{is} represents the ith consumer's implicit price per unit of characteristic s. For all consumers, the price of a good will equal the sum of its characteristics weighted by the implicit price per unit of characteristic.[14] This is precisely the content of equation (26).

If there are as many goods as characteristics (i.e., complete markets), then the implicit prices v_{is} will be the same for all consumers. This follows since in matrix form (26) can be written

$$v_i' C(q) = p' \tag{27}$$

where $v_i = (v_{i1}, \ldots, v_{iS})$.

From (24), complete markets implies the existence of an $A(q)$ such that $C(q)A(q) = I_S$. From (27) we have $v_i' C(q)A(q) = p' A(q)$, or

$$v_i' = p' A(q) \tag{28}$$

The right side of (28) is independent of i, implying implicit characteristic prices are the same for all i. If markets are not complete, the v_i will not be identical, but from (27) will lie in a subspace of dimension S-J.

In general (and in common with other equilibrium systems), the equilibrium implicit characteristic prices which satisfy (26) or (27) will depend upon y^j and q^j, the quantity and quality decisions of the firms. But the *perception* of this dependence is crucial. If markets are completely competitive, firms do not perceive their decisions affecting implicit prices v_i. Thus perfectly competitive firms perceive $\partial v_{is}/\partial q^j = \partial v_{is}/\partial y^j = 0$, for all i, j, and s. This in turn implies from (26) that firms perceive

$$\frac{\partial p^j}{\partial q^j} = \sum_s v_{is} \frac{\partial c_s^j(q^j)}{\partial q^j} \qquad j = 1, \ldots, J \tag{29}$$

$$\frac{\partial p^k}{\partial q^j} = 0 \qquad j \neq k$$

$$\frac{\partial p^j}{\partial y^j} = \frac{\partial p^k}{\partial y^j} = 0 \qquad \text{for all } j, k \tag{30}$$

Will $\partial p^j/\partial q^j$ be the same, no matter whose implicit price vector is used by the firm to compute (29)? The answer is yes, if (and only if) the spanning property is satisfied. Using (20) we have

$$\frac{\partial p^j}{\partial q^j} = \sum_s v_{is} \frac{\partial c_s^j(q^j)}{\partial q^j} = \sum_s v_{is} \sum_k c_s^k(q^k) h_k^j(q) \tag{31}$$

[14] This will not necessarily hold if corner solutions exist. see the Appendix.

$$= \sum_k \left[\sum_s v_{is} c_s^k(q^k) \right] h_k^j(q)$$
$$= \sum_k p^k h_k^j(q)$$

using (26).

The right side of (31) and hence the right side of (29) are independent of i, implying that—when spanning is satisfied—the firm will be able to compute a unique $\partial p^j/\partial q^j = \Sigma_k p^k h_k^j(q)$.

With perfectly competitive characteristic prices, (30) implies the usual competitive assumption that—for any level quality—output price will be viewed as invariant to output quantity. *But the assumption of perfectly competitive characteristic prices gives us more:* namely, a competitive response of price to a change in quality. Coupled with spanning, this response is independent of whose implicit characteristic prices are used. Therefore, the manager could in principle use his own tradeoffs between money and units of characteristics to compute $\partial p^j/\partial q^j$. Note from (23) that the firm's perceived tradeoff between price and quality coincides with what each consumer is willing to pay.

VII. Optimality of Quality Choices by Firms

We shall now show that equilibrium with quality choice will satisfy the necessary conditions (13)–(19) if (and only if, in a context made precise below) the spanning property and competitive characteristic price property are satisfied. For convenience, we regroup the equations describing equilibrium, given the spanning and competitive characteristic price assumption. We have

(32) $\sum_s U_{is} c_s^j(q^j) - \lambda_i p^j = 0$ from (10)

(33) $p^j(q^j, y^j) + \mu^j f_y^j = 0$ from (4), using (30)

(34) $-r + \mu^j f_x^j = 0$ from (5)

(35) $\sum_k p^k h_k^j y^j + \mu^j f_q^j = 0$ from (3) using (29) and (31)

(36) $f^j(q^j, y^j, x^j) = 0$ from (6)

(37) $\sum_j x^j = \bar{x} = \sum_\ell \bar{x}_{i\ell}$ from (12)

(38) $\sum_i y_i^j = y^j$ from (11)

Sufficiency is shown by demonstrating that the decisions satisfying equilibrium conditions (32)–(38) for $i = (1, 2)$ also satisfy the conditions (13)–(19), necessary for Pareto optimality.

The reader can confirm that the decisions satisfying the equilibrium (32)–(38) will also satisfy (13)–(19) with $\lambda^* = \lambda_1/\lambda_2$, $\mu^{j*} = \mu^j\lambda_1$, and $\gamma^* = -\lambda_1 r$.

If the firm's manager uses his own implicit characteristic prices for determining the $\partial p^j/\partial q^j$, we can show that spanning is a necessary as well as sufficient condition for Pareto optimality. From (3), we may substitute for $\mu^j f_q^j$ in (16), yielding the necessary condition

(39) $v_1' c_q^j y_1^j + v_2' c_q^j (y^j - y_1^j) - \frac{\partial p^j}{\partial q^j} y^j = 0$

where v_1' is the row vector with elements U_{1s}/λ_1, etc.

Without loss of generality we can assume the manager of the firm is individual 2. Using (29), (39) becomes

$$v_1' c_q^j y_1^j + v_2' c_q^j (y^j - y_1^j) - v_2' c_q^j y^j = (v_1' - v_2') c_q^j y_1^j = 0$$

But this condition can not be satisfied for arbitrary v_1 and v_2 if c_q^j is not spanned. In contrast with the situation with spanning, $v_i' c_q^j$ will not be the same for all individuals.

The necessity of perfectly competitive characteristic prices also can be shown: otherwise, rates of substitution between output and inputs, output and quality, or quantity and inputs will differ between firms and consumers.

VIII. Pareto Optimality Without Spanning

In the previous section, it was shown that spanning was a necessary condition for Pareto optimal quality decisions if we associate the firm's implicit prices with an individual's im-

plicit prices (for example, the manager). If we drop this requirement, Pareto optimality is possible when spanning is not satisfied. From (39), Pareto optimality requires

$$(40)\quad \partial p^j/\partial q^j = [v_1'(y_1^j/y^j) + v_2'(y_2^j/y^j)]c_q^j$$

where $y_2^j = (y^j - y_1^j)$ as before. More generally, for an n-person economy, we can extend (40) to show that Pareto optimality requires the firm j to act as if it had implicit prices $v_j = (v_{j1}, \ldots, v_{jS})$ such that

$$v_j = \sum_i v_i\, O_i^j$$

where $O_i^j = y_i^j/y^j$, the share of the jth good consumed by i. That is, the firm must act as if it had implicit prices which were a weighted average of implicit prices v_i, $i = 1, \ldots, I$. The weights are simply the share of the total consumption of good j consumed by person i.[15]

If the firm uses v_j, and treats these prices as invariant to its decisions y^j and q^j, the optimality conditions in Section V will be satisfied. The problem, of course, is that computation of v_j will not in general be possible, since it requires knowledge of the unobservable v_i

IX. Monopoly and Quality Choice

We have seen that, given spanning, firms will make quality and quantity decisions consistent with Pareto optimality if characteristic prices are viewed as parameters. What if these prices are perceived to depend on decisions by a firm? Can we say anything about the quality choice of a monopolist?

We shall continue to maintain the spanning hypothesis. But characteristic prices $v = (v_1, \ldots, v_S)$ are presumed to depend on the supplies of characteristic s provided by the firm. That is, $v_s = v_s(z_s)$, where $z_s = c_s(q)y$, the supply of characteristic s when the monopolist chooses quality q and output y.[16] Since by (26), $p(q, y) = \Sigma_s v_s[c_s(q)y]c_s(q)$, the profit-maximizing firm will maximize

$$\pi(q, y) = p(q, y)y - TC(q, y)$$
$$= \sum_s v_s\ [c_s(q)y]c_s(q)y - TC(q, y)$$

First-order conditions yield

$$(41)\quad \frac{\partial \pi}{\partial y} = p(q, y) + y\frac{\partial p}{\partial y} - \frac{\partial TC}{\partial y} = 0$$

$$\text{or}\quad p(q, y) + y\left\{\sum_s \frac{\partial v_s}{\partial z_s}\ [c_s(q)]^2\right\} - \frac{\partial TC}{\partial y} = 0$$

$$(42)\quad \frac{\partial \pi}{\partial q} = y\frac{\partial p}{\partial q} - \frac{\partial TC}{\partial q} = 0 \quad \text{or}$$

$$y\left[\sum_s v_s \frac{\partial c_s}{\partial q} + y\sum_s \frac{\partial v_s}{\partial z_s}\frac{\partial c_s}{\partial q}c_s(q)\right] - \frac{\partial TC}{\partial q} = 0$$

Since in (21) we identified $\Sigma_s v_s(\partial c_s/\partial q)$ with the amount every consumer would just be willing to pay per unit output for an increase in quality, we see

$$y\left[\sum_s v_s \frac{\partial c_s}{\partial q}\right]$$
$$= \text{social value of change in quality}$$

Similarly, $p(q, y)$ = social value of change in quantity; $\partial TC/\partial q$ and $\partial TC/\partial y$ represent social costs of a change in quality and quantity if other markets are competitive

From (41), we see that if $\partial v_s/\partial z_s < 0$, $p(q,y) > \partial TC/\partial y$ at the profit-maximizing output. This is the standard result that monopolists produce too little, *given* their quality level.

The sign of $y\Sigma_s(\partial v_s/\partial z_s)(\partial c_s/\partial q)c_s(q)$ in (42) can be either negative or positive. If $\partial c_s/\partial q > 0$ for all s, however—implying q unambiguously increases quality—then the term will be negative, implying

$$y\sum_s v_s \frac{\partial c_s}{\partial q} > \frac{\partial TC}{\partial q}$$

i.e., the social utility of a small increase in q exceeds its social cost. In this case, the *monopolistic firm tends to underprovide quality, given its output.*

Even assuming concavity of a social wel-

[15] Drèze derives a similar result in the context of uncertainty.

[16] For simplicity, we omit superscripts j, since we focus on a single monopolist. Of course, the equilibrium v_s depends not only on z_s, but on suppliers of all other characteristics as well. I shall ignore such interdependencies in my analysis

fare function in y and q, some care must be used in interpreting conditions $p > \partial TC/\partial y$; $y\Sigma_g v_g(\partial c_g/\partial q) > \partial TC/\partial q$. It does not necessarily follow that competition will lead both to greater output *and* to greater quality—although we cannot exclude that possibility. What we can exclude is competition leading both to lower output and lower quality. Yet it is possible that competition could lead to higher output and lower quality, or even lower output with higher quality. The correct inference as to where the monopolist "sins" depends upon complicated elasticities of several functions.[17]

X. Conclusion

While recognizing the importance of quality choice by firms, traditional economic models have been unable to examine or to explain these decisions. By using the "characteristics" approach to consumer choice, we have developed a framework for simultaneously considering quality and quantity choices by profit-maximizing firms. The characteristics approach introduces a natural metric for "distance" between goods of different qualities. Marginal analysis can then be used to examine quality choice.

Our fundamental concern was with the welfare implications of firms' quality decisions. Two properties proved essential to Pareto optimality: spanning and competitive implicit characteristics prices. The spanning property assures a single "willingness to pay" per unit consumption for a small change in quality. That is, spanning guarantees consumer unanimity with respect to the tradeoff between price change and quality change. Competitive implicit prices for characteristics, presumed invariant to firms' quality and quantity decisions, were shown to guarantee that profit-maximizing firms in equilibrium will have the same tradeoff between price and quality change as the consumers have. Spanning and competitive implicit prices were shown to be sufficient to satisfy the necessary conditions of Pareto optimality.

Competitive characteristics prices also are necessary for the optimality of quality choices. Monopolies, for example, were shown to underprovide quality, given the level of output chosen. Spanning is a necessary condition if firms are operated by managers who use the implicit prices of some individual (say, himself). If we permit firms to use weighted averages of individuals' prices, Pareto optimality is possible without spanning. In this case, the "public good" nature of quality decisions becomes evident. All consumers of a good are affected by quality changes, although differently. It was shown that Pareto optimality requires that the firm use a tradeoff between price and quality proportional to the sum of individuals' tradeoffs, weighted by their consumption of the good. The problems associated with firms implementing such a scheme are of precisely the same nature as those encountered with deciding the value of a public good.[18]

My analysis has treated the number of different quality levels (although not their location) as a constant. Welfare may be improved by creating more quality levels, just as welfare under uncertainty can be improved by a movement towards more complete markets. Of course, a greater number of markets may incur resource costs which exceed the benefits resulting from a wider selection of goods. We must await the development of a theory which endogenously explains the number of markets in existence. Such a theory will permit a final assessment of quality choices by firms in differing market environments.

[17]The nature of our results was anticipated in part by Edward Chamberlin. "The conclusion seems to be warranted that just as, for a given 'product,' price is inevitably higher under monopolistic than under pure competition, so, for a given price, 'product' is inevitably somewhat inferior. After all, these two propositions are but two aspects of a single one. If a seller could, by the large scale of production which is characteristic of pure as compared with monopolistic competition, give the same 'product' for less money, he could, similarly, give a better 'product' for the same money" (p. 99)

[18]See Paul Samuelson and related literature. Drèze and Drèze and D. de la Vallée Poussin have encountered a similar public good property of private firms' decisions in different contexts. The formal similarity between our problem and that studied by Drèze is striking.

APPENDIX

Corner Solutions

To arrive at the "unanimity condition"

$$\partial p_i^j / \partial q_j = \sum_k p^k h_k^j(q) \tag{23}$$

where the right-hand side is independent of i, we required that the first-order conditions be satisfied with equality (interior solutions) and that c_q^j be spanned by the set of securities (i.e., by the matrix $c(Q)$). If optimal consumer choice involves zero consumption of some commodities (corner solutions), spanning alone will not guarantee unanimity.

We can, however, readily modify our criterion for (23) to be satisfied for positive consumers of good j. (Note that if a consumer does not consume j, small quality changes in j will not affect his utility. We can exclude him from welfare considerations resulting from quality change in J.) Let I^j be an index set of positive consumers of j. That is, $y_i^j > 0$ if $i \in I^j$. Then the following proposition is obvious:

Let K^j be an index set of securities which span c_q^j. Then if $y_i^k > 0$ for all $i \in I^j$ and $k \in K^j$, the unanimity condition (23) *will hold for all relevant consumers* ($i \in I^j$).

Note that complete markets, which guarantee spanning, will not necessarily guarantee unanimity. Unanimity is, of course, the key aspect of Pareto optimality, and spanning should be replaced by "unanimity" in all optimality theorems if corner solutions are possible.

REFERENCES

G. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

Edward H. Chamberlin, *The Theory of Monopolistic Competition*, 6th ed., Cambridge, Mass. 1948.

Gerard Debreu, *Theory of Value*, New York 1959.

R. Dorfman and P. Steiner, "Optimal Advertising and Optimal Quality," *Amer. Econ. Rev.*, Dec. 1954, *44*, 826–36.

J. Drèze, "A Tatonnement Process for Investment Under Uncertainty," in Giorgio Szego and Karl Shell, eds., *Mathematical Methods in Investment and Finance*, Amsterdam 1972.

——— **and D. de la Vallée Poussin**, "A Tatonnement Process for Public Goods," *Rev. Econ. Studies*, Apr. 1971, *38*, 133–50.

F. Fisher, Z. Griliches, and C. Kaysen, "The Costs of Automobile Style Changes since 1949," *J. Polit. Econ.*, Oct. 1962, *70*, 433–51.

H. S. Houthakker, "Compensated Changes in Quantities and Qualities Consumed," *Rev. Econ. Stud.*, 1952, No. 3, *19*, 155–64.

K. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–57.

———, "Optimal Product Differentiation," *Amer. Econ. Rev.*, Sept. 1975, *65*, 567–85.

H. Leland, "Capital Asset Markets, Production, and Optimality: A Synthesis," tech. rep. no. 115, IMSSS, Stanford Univ., Dec. 1973.

R. Muth, "Household Production and Consumer Demand Functions," *Econometrica*, July 1966, *34*, 699–718.

R. Radner, "A Reformulation of the Ekern-Wilson Model in Terms of Arrow-Debreu," *Bell J. Econ.*, Spring 1974, *5*, 171–80.

S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 34–55.

J. Rosse, "Product Quality and Regulatory Constraints," Center Res. Econ. Growth, memo no. 137, Stanford Univ. 1972.

P. Samuelson, "The Pure Theory of Public Expenditures," *Rev. Econ. Statist.*, Nov. 1954, *36*, 387–89.

M. Spence, "Product Selection, Fixed Costs, and Monopolistic Competition," IMSSS work. pap. no. 157, Stanford Univ. 1975.

J. Sweeney, "Quality, Commodity Hierarchies, and the Housing Market," *Econometrica*, Jan. 1974, *42*, 147–67.

Residential Decentralization, Land Rents, and the Benefits of Urban Transportation Investment

By WILLIAM C. WHEATON*

In postwar America, the planning of urban transportation by federal, state, and local authorities has come to depend increasingly on economic decision making. Perhaps to a greater extent than in other areas of government investment, urban transportation has developed an elaborate planning methodology based on the principles of cost-benefit analysis (see Roger Creighton, John Meyer and Mahlon Strazheim, Herbert Mohring). In the first stage of this process, expanded transport facilities are seen to reduce the effective price (including time) of travel. In the short run this increases the number of trips, while in the long run it encourages urban decentralization and greater lengths of trips. Both forces increase aggregate travel, and the consumer surplus thus generated can be determined from an estimated demand function. With this methodology marginal cost-benefit calculations have been used both in the selection of individual projects and in the determination of aggregate investment levels.

This methodology has been criticized frequently over the years, mainly because it is a partial approach and appears to ignore the long-run repercussions of highway investment in the "adjoining" market for land and housing. The changes in rents and density that invariably follow investment have often raised the question whether benefits or costs are being created in addition to those accruing directly to highway users. In dealing with this problem, writers to date have developed quite different opinions.

Early authors, such as Robert M. Haig, suggested that land prices fully capitalized the benefits received by highway users, so that any increase in those benefits would only show up in higher land rents. It was suggested that to consider both changing rents and user savings would amount to "double counting." Mohring cast serious doubt on this proposition by arguing that while a reduction in travel costs must surely generate benefits, aggregate land payments might increase or decrease. Clearly in this situation, user benefits and changing land rents could not be equivalent. Anne Friedlaender expanded this view, arguing that changing land rents represented an additional benefit in the land market—distinct from that accruing to highway users. There was some question, however, about whether it should be counted. In the short run, such changes represent capital gains or losses, while in the longer run they are only transfers between tenant and landlord. Within the traditional benefit framework, neither makes a contribution to *GNP*.

The question of what role land rents play in benefit estimation has also been studied by those interested in the impact of nontransportation investment. Jerome Rothenberg states, for example, that the benefits of urban renewal must be measured by the aggregate change in land rents, while Robert Lind suggests that only rent changes in the immediate area are relevant for determining willingness to pay. In the most recent contribution to the debate David Pines and Yoram Weiss find that a weighted difference between rent increases in the affected area and rent decreases in other areas is a more appropriate measure. Their model, however, has several unrealistic features. Land supply is fixed, land consumption is joint between areas, and rental income is ignored. Perhaps most im-

*Assistant professor, departments of economics and urban planning, Massachusetts Institute of Technology
I would like to thank J. Rothenberg, A. Friedlaender, and several referees for their thoughtful comments and suggestions.

portant, they conclude with the view that benefits should be determined with a general equilibrium measure of income compensation. Mitchell Polinsky and Steven Shavell agree with this position, stating that the full benefits of a change in air quality must be evaluated with a spatial model, in which the general equilibrium "income equivalent" can be determined.

The most recent attempt to develop a spatial-equilibrium model of transportation investment comes from Robert Solow and William Vickrey (1971, 1973). In a series of papers, they explore the question of what optimal amount of land should be devoted to urban transportation. Increasing such land (investment) reduces congestion and spatially stimulates the demand for residential land consumption. On the cost side, however, greater land devoted to transportation restricts the supply available for residential use. The optimal tradeoff results from the solution to an extremely complicated system of equations. Even in numerical examples, it has proven difficult to extract an optimal investment rule and compare it with present practices. In principle, however, a general equilibrium model, such as Solow's, is the correct approach, and the one which will be developed here.

It is the purpose of this paper to demonstrate that much of this debate has been unnecessary. The appropriate measure of user benefits *is* equivalent to a general equilibrium "income compensation" value for highway investment. The changes in land rents and urban housing that follow highway investment need not be separately considered if the forecast of highway user demand implicitly incorporates such changes. This interpretation is not really new. It follows from the traditional literature on consumer surplus (as summarized by Arnold Harberger) and has been given added emphasis by Daniel Wisecarver's recent analysis of factor distortion. Within this latter framework, urban commuting might be viewed as a "factor" necessary for the consumption of housing and land. As long as the demand curve for the factor is a derived demand curve, its consumer surplus suffices as a measure of benefits and all other changes in final commodity markets (such as land) can be ignored.

To verify this result, the model used in this paper treats the "price" of travel as exogenous. Given this parameter land use and commuting patterns evolve from a well known model of spatial equilibrium in which aggregate rental payments are included as part of consumer income. Within this framework, transportation investment reduces the price of travel, alters the pattern of land use and commuting, and changes both income and rent payments. Acting together these alter the equilibrium level of welfare achieved in the city. The change in exogenous income necessary to compensate for this is precisely equal to the change in consumer surplus under the implicit aggregate travel demand function.

Section I of the paper develops the model and some needed intermediate results. Section II derives the general equilibrium measure of income compensation and proves its equivalence to the simple change in consumer surplus. Finally, Section III concludes with some comments on the applicability of these results to other forms of public investment.

I

In the short run, urban land use is rigid and consumer locational decisions are based on the characteristics of the standing housing stock and the resultant pattern of spatial externalities. As the time horizon lengthens, capital becomes mobile and land use is determined primarily by a long-run tradeoff between travel and residential density. The forces shaping this decision can be described either as consumer utility maximization (Richard Muth, Edwin Mills) or as rent maximization by landowners (William Alonso). Although both approaches have been shown to be equivalent (the author, 1976), the Alonso theory is mathematically more manageable and is therefore used here.

The most simple of cities is composed of N individuals with identical tastes and, for the moment, exogenous and identical incomes (y). Their utility depends on land consumption q

and a "composite" commodity x. Their incomes are divided into expenditure on x (whose price is unity), land (whose rent is r) and travel cost to a central employment district. The latter is assumed linear and equal to k times distance t. Within this environment, Alonso's theory recognizes that although consumers will live at different locations, equilibrium requires that they enjoy the same level of utility. This condition and a rewritten budget constraint are

$$(1) \qquad u = u(x, q)$$

$$(2) \qquad r = (y - x - kt)/q$$

The market process of competitive land bidding insures that landlords will eventually extract the maximum savings that consumers may enjoy—given a level of indifference u To do otherwise would be Pareto inefficient, for then a resource (land) would not be priced according to its true and highest use. The consumer variables x and q are thus determined so as to maximize (2) subject to (1). The first-order condition is (3) which together with (1) is solved for x and q, given the parameters u, t, y, and k.

$$(3) \qquad \frac{\partial u}{\partial q} \Big/ \frac{\partial u}{\partial x} = (y - x - kt)/q$$

The offered or "bid" rent of consumers is obtained by inserting the solutions for x and q into (2) and obtaining r as a function also of u, t, y, and k:

$$(4) \qquad \begin{array}{c} r(u, t, y, k) \\ x(u, t, y, k) \\ q(u, t, y, k) \end{array}$$

Since r is maximized with respect to x and q for any value of the parameters, the envelope theorem is used to determine the influence of these. In particular:

$$(5) \quad \frac{\partial r}{\partial u} = -1 \Big/ \frac{\partial u}{\partial x} q < 0 \qquad \frac{\partial r}{\partial y} = 1/q > 0$$

$$\frac{\partial r}{\partial k} = -t/q < 0 \qquad \frac{\partial r}{\partial t} = -k/q < 0$$

A general equilibrium requires an indifference level for consumers which exactly balances the supply of land with that demanded at the same u. Land supply reaches from the city center to a location b, where consumer bid rents equal some opportunity cost of land s. Demand and supply are balanced when the holding capacity of land up to this point equals the long-run population to be housed N. These equilibrium conditions, (6)–(7), are solved simultaneously for the boundary b and level of indifference u.[1]

$$(6) \qquad r(u, b, y, k) = s$$

$$(7) \qquad 2\pi \int_0^b t/q(u, t, y, k)\, dt = N$$

By incorporating (5) and integrating by parts, it is possible to rewrite (7) as (7′) below:

$$(7') \qquad bs - \int_0^b r(u, t, y, k)\, dt = -kN/2\pi$$

The solution to the system of equations (6)–(7), or alternatively (6)–(7′), represents a mapping from (y, k) to (u) which may be represented by (8):

$$(8) \qquad u = U(y, k)$$

While the model so far treats income as exogenous, the existence of rent payments clearly suggests it should be endogenous. Rental income arises from both the rural opportunity use of land (valued at s) and urban rent payments (r). If the total land area in the country is A, income from rural users will be $s(A - \pi b^2)$, while that from urban users will equal the integral of r from the city center to its boundary at b. Total household income will then be the sum of exogenous nonwage receipts (y_0) plus an equal share of aggregate rents (R). This is elaborated in expression (9).

$$(9) \quad y = y_0 + \frac{2\pi}{N} \int_0^b r(u, t, y, k) t\, dt + s(A - \pi b^2)/N = y_0 + R(u, y, k)/N$$

The broader system of equations (8)–(9) is now solved for both y and u as function of y_0

[1]This model assumes that the city is "closed," that is, that welfare levels are endogenous An alternative assumption is that a large nonurban sector sets an exogenous level of welfare, and migration equalizes urban-rural utility levels. For developing nations this latter model may be useful, but for Western industrial countries, clearly the closed city is more realistic.

and k. As k is changed (through transportation investment), the exogenous income necessary to fully compensate is

$$\frac{dy_0}{dk} = -\frac{du}{dk}\Big/\frac{du}{dy_0} \tag{10}$$

The general equilibrium effect of k on u is determined by the following total differentiation of (8)–(9):

$$\frac{du}{dk} = \frac{\partial U}{\partial y}\frac{dy}{dk} + \frac{\partial U}{\partial k}$$

$$N\frac{dy}{dk} = \frac{\partial R}{\partial y}\frac{dy}{dk} + \frac{\partial R}{\partial k} + \frac{\partial R}{\partial u}\frac{du}{dk}$$

This solves to:

$$\frac{du}{dk} = \frac{N\frac{\partial U}{\partial k} - \frac{\partial R}{\partial y}\frac{\partial U}{\partial k} + \frac{\partial U}{\partial y}\frac{\partial R}{\partial k}}{\left(N - \frac{\partial R}{\partial y} - \frac{\partial R}{\partial u}\frac{\partial U}{\partial y}\right)} \tag{11}$$

The general equilibrium effect of y_0 on u is determined by similar differentiation to be:

$$N\frac{dy}{dy_0} = N + \frac{\partial R}{\partial y}\frac{dy}{dy_0} + \frac{\partial R}{\partial u}\frac{du}{dy_0} \tag{12}$$

$$\frac{du}{dy_0} = \frac{\partial U}{\partial y}\frac{dy}{dy_0}$$

$$= N\frac{\partial U}{\partial y}\Big/\left(N - \frac{\partial R}{\partial y} - \frac{\partial R}{\partial u}\frac{\partial U}{\partial y}\right)$$

Combining these results, the general equilibrium income compensation for a change in k is the ratio of (11)/(12), or:

$$\frac{dy_0}{dk} = \left(-\frac{\partial U}{\partial k}\Big/\frac{\partial U}{\partial y}\right) - \frac{1}{N}\left[\frac{\partial R}{\partial k} + \frac{\partial R}{\partial y}\left(-\frac{\partial U}{\partial k}\Big/\frac{\partial U}{\partial y}\right)\right] \tag{13}$$

The first term on the right side of (13) is the direct income compensation necessary in the partial model (8) where all income is exogenous The second term (within brackets) is the net change in per capita rental income that results from both the initial change in k and the offsetting income compensation. The difference between these terms is the total income necessary to keep utility constant—given that changing rental income helps to offset the direct compensation. It is important to remember that this change in rental income is computed assuming that the income compensation is actually paid. Any measured changes in rent will be based on altered utility levels, and hence are not equivalent to the second term in expression (13). Of course, a full analysis of (13) requires a more thorough evaluation of the four derivatives $\partial U/\partial k$, $\partial U/\partial y$, $\partial R/\partial y$, $\partial R/\partial k$.

II

The derivatives of (8), $\partial U/\partial y$ and $\partial U/\partial k$ can be found by totally differentiating the system of equations (6) and (7′) to which (8) is the solution. Considering first the impact of travel costs, (7′) is differentiated to yield:[2]

$$s\frac{db}{dk} - \int_0^b \left(\frac{\partial r}{\partial u}\frac{du}{dk} + \frac{\partial r}{\partial k}\right) dt - r_b\frac{db}{dk} = -N/2\pi$$

Since rent at $b(r_b)$ must equal s, two terms cancel and du/dk can be determined without recourse to differentiating (6).

$$\frac{\partial U}{\partial k} \equiv \frac{du}{dk} = \left[N/2\pi - \int_0^b \frac{\partial r}{\partial k}dt\right]\Big/\int_0^b \frac{\partial r}{\partial u}dt \tag{14}$$

Differentiating (7′) again, this time with respect to y yields:

$$s\frac{db}{dy} - \int_0^b \left(\frac{\partial r}{\partial u}\frac{du}{dy} + \frac{\partial r}{\partial y}\right) dt - r_b\frac{db}{dy} = 0$$

which solves to:

$$\frac{\partial U}{\partial y} \equiv \frac{du}{dy} = -\int_0^b \frac{\partial r}{\partial y}dt\Big/\int_0^b \frac{\partial r}{\partial u}dt \tag{15}$$

The ratio equals:

$$\frac{\partial U}{\partial k}\Big/\frac{\partial U}{\partial y} = \left[N/2\pi - \int_0^b \frac{\partial r}{\partial k}dt\right]\Big/ - \int_0^b \frac{\partial r}{\partial y}dt \tag{16}$$

Remembering (from (5)) that $\partial r/\partial k$ equals

[2]Letter subscripts such as r_b refer to the location at which a particular function (r) is being evaluated.

$-t/q$ and $\partial r/\partial y$ equals $1/q$, expression (16) together with (7) reduces to:

$$(17) \qquad \frac{\partial U}{\partial k} \Big/ \frac{\partial U}{\partial y} = -N/\pi \int_0^b 1/q\, dt$$

A simpler version, without the integral, is obtained when it is recalled that according to (5):[3]

$$\pi \int_0^b 1/q\, dt = -\pi \int_0^b \frac{\partial r}{\partial t} \Big/ k\, dt = \pi \frac{(r_0 - s)}{k}$$

and hence:

$$(18) \qquad \frac{\partial U}{\partial k} \Big/ \frac{\partial U}{\partial y} = -Nk/\pi(r_0 - s)$$

The derivatives $\partial R/\partial y$, $\partial R/\partial k$ are found by partially differentiating the single equation (9). Considering y first:

$$(19) \quad \frac{\partial R}{\partial y} = 2\pi \int_0^b \frac{\partial r}{\partial y} t\, dt + 2\pi r_b \frac{\partial b}{\partial y} - 2\pi s b \frac{\partial b}{\partial y} = 2\pi \int_0^b t/q\, dt = N$$

The latter steps follow from incorporating (5) and (7). Similarly, the partial impact of k is:

$$(20) \qquad \frac{\partial R}{\partial k} = 2\pi \int_0^b \frac{\partial r}{\partial k} t\, dt + 2\pi r_b \frac{\partial b}{\partial k} = -2\pi \int_0^b t^2/q\, dt$$

Combining (18), (19), and (20) into (13) the final expression for the general equilibrium income compensation is:

$$\frac{dy_0}{dk} = \frac{Nk}{\pi(r_0 - s)} - \frac{1}{N}\left[-2\pi \int_0^b t^2/q\, dt + \frac{N^2 k}{\pi(r_0 - s)}\right] = +\frac{2\pi}{N} \int_0^b t^2/q\, dt$$

[3]The assumption is made here that the form of the r function leads to central rents r_0 which are bounded

The aggregate benefit of a change in k is simply N times the individual income compensations, or:

$$(21) \qquad \frac{dy_0}{dk_{Aggregate}} = 2\pi \int_0^b t^2/q\, dt$$

This expression is easily interpreted when it is remembered that at each location t there are $2\pi t/q$ individuals who are all commuting t miles. Integrating the product of these terms from city center to boundary yields the aggregate miles traveled as a consequence of the existing pattern of residential density. Of course, this density gradient, and hence aggregate travel, is dependent on the parameters y_0 and k. As the price of travel falls, or income rises, the city boundary expands, residential density is lowered, and aggregate travel increases (see the author, 1974). The right side of expression (21), therefore, is an aggregate travel demand function in income and the price of travel. The marginal change in consumer surplus from an alteration in price is simply the level of consumption of that commodity whose price was changed. Thus expression (21) equates the general spatial equilibrium "income compensation" value of a marginal highway investment with the marginal change in consumer surplus under the derived demand function for travel. All of the changes in the housing and land market that accompany highway investment can be completely ignored in benefit calculations if highway demand is adequately forecasted.

III

It would be tempting to conclude that the results of the analysis pertain to all government investments that have impacts in the land market. In this case, the persistent problem of measuring and evaluating these impacts might be avoided. Unfortunately, this extrapolation is premature. To begin with, many types of investment have direct influences on land and housing, not just indirect ones through the alteration of some other market price. A nuclear generating plant, for example, has its direct impact in the market for electricity. While it is doubtful

that this in turn induces any change in land rents, the externalities from the plant might indeed affect the surrounding land market. This latter influence seems more "direct" and hence not the same as the induced change from transportation investment. Most likely this should be separately counted in some way along with the benefits of cheaper electricity. Urban renewal presents a similar case, in which the proposed investment policy (cheaper land development) will have a direct impact in the land market rather than only a secondary "induced" change. Clearly, it seems important to characterize the type of influence that the investment project has on the land market before considering whether and how to evaluate it.

A second problem with the present analysis is that it ignores the externalities inherent in highway usage. To avoid the complexity of Solow's model, this paper assumed an exogenous price for commuting. In fact, the price faced by transportation users is an endogenous function of the extent of their usage In the absence of appropriate congestion "tolls," this price will not equal social cost, and so the market for transportation will contain a distortion. As Harberger suggests, the correct benefit measure for investment may be different in the presence of such distortion Without congestion or "peak" pricing, the benefits of urban highway investment may involve a more complicated assessment of demand than that conducted in this paper

REFERENCES

William Alonso, *Location and Land Use*, Cambridge 1964.

Roger Creighton, *Urban Transportation Planning*, Urbana 1970.

Anne F. Friedlaender, *The Interstate Highway System*, Amsterdam 1965.

Robert M. Haig, *Regional Survey of New York and Environs*, Vol 1, New York 1927.

A. C. Harberger, "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *J. Econ. Lit.*, Sept. 1971, *9*, 785–97.

R. C. Lind, "Spatial Equilibrium, the Theory of Rents and the Measurement of Benefits from Public Programs," *Quart. J. Econ.*, May 1973, *87*, 188–207.

John Meyer and Mahlon Strazheim, *Pricing and Project Evaluation Techniques*, New York 1971.

Edwin S. Mills, *Studies in the Structure of the Urban Economy*, Washington 1972.

Herbert Mohring, *Highway Benefits, An Analytic Framework*, Chicago 1962

Richard Muth, *Cities and Housing*, Chicago 1969.

D. Pines and Y. Weiss, "Land Improvement Projects and Land Values," *J. Urban Econ.*, Jan. 1976, *3*, 1–13.

M. Polinsky and S. Shavell, "The Air Pollution and Property Value Debate," *Rev. Econ. Statist.*, Feb. 1975, *57*, 100–04.

J. Rothenberg, "Urban Renewal Programs," in Robert Dorfman, ed., *Measuring the Benefits of Government Investment*, Washington 1965.

R. M. Solow and W. Vickrey, "Land Use in a Long Narrow City," *J. Econ. Theory*, Dec. 1971, *3*, 430–47.

——— **and** ———, "Congestion Cost and the Use of Land for Streets," *Bell J. Econ.*, Autumn 1973, *4*, 601–18.

A. A. Walters, "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, Oct. 1961, *29*, 676–99.

W. C. Wheaton, "A Bid Rent Approach to Housing Demand," *J. Urban Econ.*, forthcoming, Apr. 1977.

———, "A Comparative Static Analysis of Urban Spatial Structure," *J. Econ. Theory*, Oct. 1974, 9, 223–37.

D. Wisecarver, "The Social Costs of Input Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359–72.

A Bid-Rent Analysis of Housing Market Discrimination

By GEORGE C. GALSTER*

The elimination of racial discrimination has long been a dominant American social concern. Laudably, economists have contributed many studies which have attempted to identify and quantify such discrimination, particularly in the area of the housing market. The research reported here, while following in this tradition, employs a new approach in attempting to discover not only the magnitude of housing price discrimination, but how its burden is incident upon different types of nonwhite households. A model of urban housing markets will be developed from the bid-rent theory which allows one to isolate empirically the distinct contributions to interracial housing price differentials made by variations in households' preferences, incomes, and housing packages versus those made by discriminatory actions. Bid-rent functions will be econometrically estimated for individual household observations stratified into groups of comparable age, family size, education, socioeconomic class, and race. These functions will be used to estimate what the various nonwhite strata would be willing to bid for typical white-occupied units. The divergencies between such bids and prices actually paid by whites will provide a measure of the existence and magnitude of discrimination confronting nonwhite groups.

I. Review of Existing Studies

While this research is similar in some respect to numerous recent econometric investigations of housing discrimination, it utilizes a distinctive methodology in terms of theoretical underpinning and empirical specification. A brief review of the two main strands of existing econometric specifications reveals that neither can conclusively identify housing discrimination without recourse to arbitrary and often implausible assumptions concerning household preferences for neighborhood racial composition and the unique sociopsychological atmosphere generated by the ghetto environment.[1]

One set of studies has explained the variation in housing prices and rents attributable to differences in the components comprising the dwelling packages by use of a "hedonic-index" form of equation estimated over pooled samples of both races. The use of a discrete "racial" dummy variable (for example, "race of household head" as in A. T. King and Peter Mieszkowski and in Edgar Olsen, or "tract in nonwhite submarket" as in Charles Daniels) or some continuous measure of neighborhood racial composition (for example, "percentage white in tract" as in Ronald Ridker and John Henning, John Kain and John Quigley (1970), and Daniels, or "percentage nonwhite in surrounding ring of blocks" as in Martin Bailey) supposedly measures the discriminatory effects.

The problem associated with the use of only *one* such racial variable in the regression is that, given the high degree of residential segregation, it not only identifies different *groups demanding* housing but also proxies for a *component supplied* by the housing package which may have intrinsic value to either (or both) group. Two such possible race-related components of the package must be considered. The first assumes a

*Assistant professor of economics and chairman, urban studies, The College of Wooster. I wish to acknowledge gratefully the helpful thoughts concerning this research contributed by Benjamin Berry, Robert Engle, Franklin Fisher, John Naylor, Gene Pollock, Lester Thurow, William Wheaton, and an anonymous referee, while retaining full responsibility for any flaws in the analysis. Special thanks are also due John Kain and John Quigley, who generously shared their data with me for the purpose of this research.

[1] A more detailed critique of existing methodologies is found in my 1974 and 1976 papers.

"taste for segregation" (integration), wherein people may be willing to pay a premium to live in units near households of their own race (opposite race). This attribute of "neighborhood racial composition" is usually conceived of in continuous terms—its strength varies continuously with the racial proportions of proximate neighborhoods.

The second potential race-related component of a housing package which has heretofore been overlooked in the economic literature may be called "ghetto environment." The sociological literature is rife with examples (see, for instance, Lee Rainwater, ch. 13; Daniel Moynihan and Nathan Glazer, ch. 2; Kenneth Clark; LeRoi Jones; Ulf Hannerz) of how the heart of the ghetto holds unique attributes in the eyes of nonwhites which are not present in racially integrated areas. It is unclear, however, whether the net effect of these attributes is positive or negative. On the positive side, the ghetto provides the citadel of nonwhite cultural, spiritual, and recreational activity—an area where a sense of "belongingness" or "black pride" is engendered. On the negative side, the ghetto has been characterized as a pathological concentration of deprivation, frustration, and anomie; a community that is weak, disorganized, and unable to provide constructive support or social control over its members. Regardless of which effect dominates, the ghetto environment characteristic can best be modeled in discontinuous terms—units located inside the distinctly demarcated ghetto share it to the same degree while those outside don't possess it at all.[2]

With these two potential factors in mind the sense of the above criticism should be transparent. Variables attempting to identify the race of the household in a regression not stratified by race may actually be proxying for an additional housing attribute contributing to the unit's value. For example, a positive coefficient for a nonwhite household head dummy may signify either a discriminatory markup or the positive value nonwhites place in "belonging to the ghetto community"; an attribute which, due to existing residential segregation, is usually associated with the typical nonwhite head but *not* the typical white head. Analogously, a negative coefficient for a percentage white in tract variable could mean either whites receive discounts, or nonwhites have a relatively stronger preference for living in predominantly nonwhite tracts than do whites in white tracts. Of course, the confusion can work in the opposite direction so as to obscure discrimination which may actually exist. The absence of interracial price differentials cannot rule out the possibility of discrimination if, for instance, it is working to offset the discounts generated by predominantly negative attributes characterizing ghetto location. Only after making the implausible assumption that neither neighborhood racial composition nor the environment of the ghetto are arguments in household utility functions do these ambiguities disappear.

King and Mieszkowski tried two additional specifications (based on a pooled sample of households of both races) in an attempt to avoid the above problem. They first used dummy variables for race of household head and percentage nonwhite on block in the same hedonic equation. Unfortunately, the ghetto environment component was overlooked. If this factor was, in fact, operative and highly correlated with the two above explanatory variables, its exclusion from the equation would bias the estimated coefficients of these included variables. The significance of this bias can only be dismissed by assuming ghetto residents placed no value on their environment, or that this factor was uncorrelated with other components. What's more, the variable for percentage nonwhite on block is a poor control for the taste of each race for racial composition in a nonstratified sample, unless one assumes that both races evaluate racial composition in the same degree (an assumption con-

[2]It should be noted that Kain and Quigley and King and Mieszkowski both try to control for "environmental" factors using continuous variables like quality of proximate dwellings, local school achievement, crime rates, etc. While the model reported in this paper also controls for these factors, there exists persuasive evidence for also including a discontinuous ghetto dummy variable proxying for the unique sociopsychological ghetto environment attributes.

trary to King and Mieszkowski's conclusions). Their second specification involved multiple dummy variables simultaneously delineating both location in the housing market in terms of three broad ranges of neighborhood racial composition and the race of the renter. Their claim that nonwhites pay discriminatory premiums in racially mixed "boundary" areas is mitigated, however, by the fact that their specification of this region still allowed variation of 3–60 percent nonwhite occupancy in adjacent blocks. As King indicated in personal correspondence the distribution of races across this boundary region was such that the average white renter lived in a neighborhood having a significantly lower proportion of nonwhites than the average nonwhite renter. Thus, this multiple dummy technique does not provide complete standardization of the neighborhood racial composition component of the housing package. Only by assuming nonwhites have no tastes for segregation can the above potential caveat be skirted, yet this would again be contrary to King and Mieszkowski's conclusions.

The other strand of research has utilized sample stratification in its specifications. Kain and Quigley (1970, 1974) also estimated for ghetto and nonghetto parcels separate hedonic equations containing numerous characteristics of the housing package, including a percentage white in tract variable. The coefficients of these stratified models were then applied to the mean values of the explanatory variables for units in both racial submarkets to derive price differentials.

Unfortunately, stratification by ghetto and nonghetto areas does not negate the fact that the ghetto not only may identify a housing submarket occupied by nonwhites but also the aforementioned environmental characteristic possessed uniquely by ghetto units. If this characteristic is highly valued by ghetto residents its price will implicitly be included in the estimated constant term for the ghetto stratum hedonic regression. Thus, even if the other coefficients are unbiased, the aforementioned ghetto/nonghetto rent simulations will falsely indicate a markup. Once again, one cannot be sure to avoid this potential confusion unless one assumes the ghetto environment component has no independent effect on rents.

Victoria Lapham used a similar stratification but included only those housing components in the hedonic index which were common to both races (excluding neighborhood racial composition and ghetto environment) and then tested for equality of coefficients (implicit component prices). Since this specification intentionally excludes race-related independent variables presumed to affect the package price, the specter of coefficient bias again arises. While Lapham was refreshingly forthright in her recognition of this shortcoming, nevertheless, it does not reduce its significance, precisely because the variables intentionally excluded (neighborhood racial composition and ghetto environment) are likely to be highly correlated (but, undoubtedly, to a different degree in each subsample) with some of the included variables. As before, the reliability of the approach can only be assured by the assumption of indifference to neighborhood racial composition or ghetto location.

Mahlon Straszheim (1974) employed a more sophisticated stratification technique. Straszheim estimated for nonwhites of different life cycle categories demand functions for various physical components of the housing package based on their income, prices of different "benchmark" structure types, and a ghetto submarket dummy variable. He then used these functions to compute the contribution discrimination made to the interracial difference in the consumption of the attribute in question by lowering nonwhite price-income ratios to the white level and eliminating the ghetto submarket dummy, thereby estimating an expected nonwhite consumption level. Unfortunately, his estimation of prices of benchmark units which were so crucial for the demand equations utilized a hedonic index which apparently overlooked the possibility of price variations due to different preferences for neighborhood racial composition and/or the ghetto environment. If, for instance, a certain nonwhite life cycle group placed some nonzero value on this latter factor their price-

income ratios would be higher, *ceteris paribus*, in the ghetto since unique features were present there and not in the white submarket. The upshot is that Straszheim's methodology fails to control for several possible interracial differences in preferences for race-related housing components —differences which, unless assumed away, may seriously bias estimates of the extent of hypothesized discriminatory practices.

The research reported here avoids the difficulties encountered by these existing specifications not only by more fully controlling for differences in the housing package components of neighborhood racial composition and ghetto environment, but also by conducting statistical estimations over various racially stratified subsamples of households selected to have comparable preferences for housing. Specifically, implicit bid-rent functions will be estimated for individual household observations stratified by age, family size, education, income, and race. In each function variables proxying for neighborhood racial composition and ghetto environment will appear, thus a given racial stratum's explicit evaluation of these components will be estimated. These functions, in turn, will be utilized to estimate what various nonwhite strata would hypothetically be willing to bid for benchmark white-occupied dwellings. Differences between these hypothetical bids and actual white prices indicate, then, discriminatory practices, controlling for the two race-related housing components.

The next section describes the theoretical foundation for this new approach. Following sections outline the specification in more detail and present empirical results.

II. Theoretical Framework

The analysis of housing discrimination in this paper is founded upon the bid-rent theory of urban land pricing originally presented by William Alonso. The theory considers the pricing mechanism by which vacant land surrounding some employment center is allocated to different households comprising the urban labor force. Each household formulates a "bid-rent function" showing the set of maximum per acre prices it would be willing to pay for acreage at each distance from the center while remaining at some arbitrary level of utility. These bids are a function of the household's preferences for land, travel time, and other consumer goods, its income, and the per mile, out-of-pocket transportation costs. Households compete for sites in accordance with their bid-rent functions until, in competitive equilibrium, a rent gradient is established such that: a) All households are allocated some parcel; b) Rent paid by the most distant household equals the nonurban opportunity cost of land; c) All households of identical incomes and preferences have equal levels of welfare, regardless of where they locate or what rent they pay; d) No household, regardless of income or preferences, can outbid another for a parcel while remaining at the same level of welfare.

This traditional bid-rent theory is readily adaptable to an analysis of rents in a developed city with a given array of housing structures located on parcels of given size and accessibility in given neighborhoods. The bid of a household with particular preferences and income is now not a continuous function of distance to work and lot size, but rather is defined for each discrete parcel. It thus is a "function" of the parcel's size and accessibility, the size and quality of unit located on it, neighborhood and pollution conditions, public service availability, race-related attributes, etc.

This modified "discrete" bid-rent model owes its origins to the work of Britton Harris, Josef Nathanson, and Louis Rosenberg, and of William Wheaton, and may be expressed as follows. A household faces a budget constraint:

$$B = Y - k(t) - Z \tag{1}$$

where B is the bid or total annual expenditure on the housing package, Y is annual household income, $k(t)$ is the annual out-of-pocket transportation costs associated with a particular unit's distance/travel time from work t, Z is annual expenditure on all nonhousing consumption.

The household's preferences are given by its utility function:

$$(2) \qquad u = u^*(Z, [q_i], t)$$

where $[q_i]$ is the n-vector of housing package components (rooms, age, quality, neighborhood, etc.), and u is the level of utility.

The household's bid-rent function is derived by maximizing (1) subject to (2). For each particular parcel in the set over which B is defined the only choice variable is Z since $[q_i]$ is given in the short run. Thus (2) may be solved for Z yielding the inverse function U and substituted into (1) yielding

$$(3) \qquad B = Y - k(t) - U(u, [q_i], t)$$

Equation (3) may be further developed by specifying a particular functional form for u^*. There is, unfortunately, no widely accepted form which is felt adequate to capture household preferences. It does seem reasonable, however, to posit utility functions which satisfy certain minimal criteria. They should, for example, generate convex indifference surfaces consistent with common presumptions about declining marginal rates of substitution. They should not generate indifference curves in Z-q_i space which intersect the q_i axis since that would imply a finite amount of q_i could compensate for having no other consumption. Finally, for reasons peculiar to this particular study, functions are chosen which yield bid-rent functions estimable by ordinary least squares (*OLS*) regression techniques.[3]

In light of these criteria, four utility functions are considered: Cobb-Douglas (*CD*), generalized Constant Elasticity of Substitution (*CES*), generalized power (*PWR*), and modified exponential (*EXP*):[4]

$$(4) \qquad u = Zt^{\theta}\prod_{i}^{n} q_i^{\phi_i} \qquad \theta, \phi_i > 0 \qquad (CD)$$

$$(5) \qquad u = \alpha Z^{-1} + \theta t^{-1} + \sum_{i}^{n} \phi_i q_i^{-1} \qquad \alpha, \theta, \phi_i < 0 \qquad (CES)$$

$$(6) \qquad u = \alpha Z + \theta t^{-1} + \sum_{i}^{n} \phi_i q_i^{-1} \qquad \theta, \phi_i < 0; \alpha > 0 \qquad (PWR)$$

$$(7) \qquad u = Z \cdot exp\left\{\theta t^{-1} + \sum_{i}^{n} \phi_i q_i^{-1}\right\} \qquad \theta, \phi_i < 0 \qquad (EXP)$$

Solving these functions for Z, substituting the result into (1), and rearranging the resultant form (3), we get a set of implicit bid-rent functions possessing the property that the $[q_i]$ parameters of the utility functions appear as coefficients in linear (sometimes in the *log*) equations:[5]

$$(8) \qquad ln(Y - k(t) - B) = ln\, u - \theta\, ln\, t - \sum_{i}^{n} \phi_i\, ln\, q_i \qquad (CD)$$

$$(9) \qquad (Y - k(t) - B)^{-1} = u/\alpha - (\theta/\alpha)t^{-1} - \sum_{i}^{n} (\phi_i/\alpha) q_i^{-1} \qquad (CES)$$

$$(10) \qquad Y - k(t) - B = u/\alpha - (\theta/\alpha)t^{-1} - \sum_{i}^{n} (\phi_i/\alpha) q_i^{-1} \qquad (PWR)$$

$$(11) \qquad ln(Y - k(t) - B) = ln\, u - \theta t^{-1} - \sum_{i}^{n} \phi_i q_i^{-1} \qquad (EXP)$$

Equations (8)–(11) provide the theoretical basis for the statistical estimation of implicit bid-rent functions (with the addition to each equation of a stochastic error term possessing the usual properties). For a given stratum of households with common incomes and utility functions, Y and u will be constants. Therefore, one can analyze the variations in t, $k(t)$, B, and $[q_i]$ to distill the coefficients of (8)–(11), and thus

[3]This criterion was mandated by strata sample sizes which were too small to permit estimation of two parameters for each functional argument as in non-linear estimation techniques.

[4]Note while the *CES*, *PWR*, and *EXP* forms are very similar, they yield different marginal rates of substitution in Z-q_i space and hence represent distinct utility functions. A proof that these functions satisfy the aforementioned criteria is given in the author (1974).

[5]The fact that the coefficients are not the parameters directly, but rather their ratio, is not worrisome since utility is invariant under monotonic transformation.

TABLE 1—DEFINITION OF HOUSEHOLD STRATA ANALYZED

Code	Age of Head	Children	Income	Observations
White Households				
1	under 31	0–2	$4 5–12,000	44
2	31–55	0–2	$0–5,999	30
3	31–55	0–2	$6–15,000	92
4	31–55	3+	$6–15,000	29
5	56+	0	$0–2,999	66
6	56+	0	$3–14,000	103
Nonwhite Households				
1	31–55	0–2	$0–5,999	47
2	31–55	0–2	$6–15,000	41
3	56+	0	$0–2,999	64
4	56+	0	$3–14,000	40

Note. None of the above strata contained college-educated heads.

the desired utility function parameters. From that point the parameters can be easily manipulated to obtain the explicit bid-rent function (3).

III. Specification of Household Strata

Clearly, the challenge faced when employing the foregoing bid-rent model is the specification of household groups who do in fact possess common incomes and preferences—both in terms of functional form and parameter values. As for isolating common utility functions, the strategy employed was to stratify by two general categories which were thought to capture the most important preference-determinants but were not so narrowly defined as to create unacceptably small sample sizes. The two chosen categories were stage in household life cycle and socioeconomic class.[6]

Operationally, life cycle stage was captured in the following manner. Observations were first stratified by age of head: under 31; 31–55; over 55 years. These strata were in turn subdivided by family size: under three/three and over children for the 0–30 and 31–55 age groups; and zero/nonzero children for 55+ age households. Socioeconomic class was proxied for by a further bifurcation by education (no college/at least some college), and trifurcation by lower, middle, and upper income groups: $0–4,999/$5–12,000/over $12,000 for 0–30 age; $0–5,999/$6–15,000/over $15,000 for 31–55 age; $0–2,999/$3–14,000/over $14,000 for 55+ age household categories.[7] Race of household was, of course, the final stratification criterion for each of the above cells. The actual strata used for the analysis due to their adequate sample sizes are listed in Table 1 along with their code numbers which will be used in succeeding references.

Unfortunately, even assuming the foregoing stratification succeeded in standardizing preferences it was impossible to compare households of identical utility levels. The need to maintain adequate sample sizes forced the use of strata encompassing a range of incomes with the concomitant assumption that u in (8)–(11) was some simple function of income within each stratum. Furthermore, normal market frictions like moving and information costs mean that otherwise identical households may have been at slightly different u levels because their B varied

[6]The crucial importance of life cycle category has been cited by John Lansing and Leslie Kish, Beverly Duncan and Philip Hauser, and Straszheim (1973). The independent preference-shaping power wielded by socioeconomic class has been claimed by Herbert Gans, David Birch et al., and Chester Rapkin and William Grigsby

[7]These income categories were chosen so that about 20 percent of the income distribution estimated for St Louis in 1967 for that age category was isolated in each tail. Note this specification does not imply that as income changes, so do preferences Income here is used to proxy for the broader taste-affecting category of socioeconomic status

from its true equilibrium level by some random amount.[8] To correct for this latter possibility a dummy variable (*NMOVE*) was utilized to indicate that the household had not moved in over 10 years, and perhaps had not readjusted to the current housing opportunity set.[9] In sum, u in (8)–(11) was formulated as:

$$(12) \qquad u = \Psi Y^{\rho} e^{\gamma NMOVE}$$

$$\Psi, \rho > 0; \; \gamma < 0 \text{ if } u \text{ defined by } CD \text{ or } EXP$$

$$(13) \qquad u = \Psi + \rho Y + \gamma NMOVE$$

$$\rho < 0, \; \gamma > 0 \text{ if } CES; \; \rho > 0, \; \gamma < 0 \text{ if } PWR$$

IV. Estimation of Bid-Rent Functions

The foregoing bid-rent specification was estimated via multiple regression techniques using data gathered from approximately 1100 randomly sampled individual households in the central city of St. Louis, Missouri, during 1967. These data have already been utilized in several important studies by Kain and Quigley (1970, 1972, 1974), and a detailed explanation of sample description and methodology may be found in their earlier publications. This data base has a plethora of individual dwelling unit characteristics as well as socioeconomic and demographic information concerning the occupying household.

The $[q_i]$ housing components utilized were as follows. The quantitative attributes of the dwelling were summarized by *AREA*, *AGE*, and *PARCL*—unit gross floor area in square feet, structure age in years, and parcel yard area attached to structure in which unit located, respectively. Three qualitative components distilled from a host of quality indexes via principle components analysis were used: structural quality and condition of the unit interior (*QUNIT*), aesthetic quality of residential environment (*QRENV*), and quality of adjacent structures (*QADJS*).[10] The quality of local public services was proxied for by the number of felonies in the police enumeration grid encompassing the unit (*CRIME*), and an index of physical problems or defects in the local school building (*SCHOLP*). Neighborhood racial composition was captured by the percentage of white households in the census tract encompassing the unit (*PCTWT*). Finally, a dummy variable for any tracts which were greater than 95 percent nonwhite (*GHETTO*) was included to test for any unique sociopsychological attributes of the ghetto environment.

The oft-mentioned need to maintain adequate stratum sample sizes forced the pooling of owner and renter occupant households. The approach in estimating B for each observation was to include only those annual expenditures which were intrinsically related to the housing structure and independent of the particular tastes, incomes, and family size of the occupants. Thus, for owners, B was computed as the sum of property tax payments, maintenance expenditures, opportunity cost of equity capital, and bills for water and heating For renters, B was annual contract rent plus annual costs for stove rental, water, and heat (if these were *excluded* from stated rent), less annual costs for refrigerator and furniture rentals and electricity bills (if these are *included* in stated rent). This procedure yielded B for owners which were 10–15 percent of market value, i.e., about 1 percent per month, which is, of course, consistent with the widely used 100:1 ratio converting monthly rents to market values.

Finally, Y was directly available from the data and, although it represented only current and not permanent income of the household, it was assumed no serious bias was produced. Annual $k(t)$ was estimated from data on work travel times (t) and modes.[11]

For every stratum, four *OLS* regressions were

[8]Another factor leading to utility variations within strata is the existence of multiple employment centers, see, for example, Leon Moses

[9]*NMOVE* is clearly a crude proxy for frictional disequilibrium since a resident can adjust housing consumption not only by moving but also by altering the quality of the existing unit.

[10]The exact components comprising these quality indices and their factor loadings are found in Kain and Quigley (1970).

[11]The detailed procedures used for estimating all these factors are described in the author (1974).

TABLE 2—INCREMENTAL ANNUAL DOLLAR VALUE OF HOUSING COMPONENTS FOR NONWHITES

Components	Nonwhite Strata 1 (*EXP*)	2 (*CES*)	3 (*PWR*)	4 (*CD*)
AREA	$ 88	$−15.40	$ 2.99	$ −14.99
	(19.7)	(21.4)	(11.0)	(25.8)
AGE	−8.86	−17.04	−3.29	−11.79
	(4.8)[a]	(3.5)[a]	(2.7)[c]	(7.0)[b]
PARCL	53	2.68	0.00	4.62
	(0.5)[c]	(2.3)[c]	(0.0)	(3.6)[c]
QUNIT	−.71	−4.74	−2.01	14.39
	(7.8)	(7.0)	(3.4)	(10.6)[b]
QRENV	2.74	−8.03	7.28	11.61
	(7.3)	(11.8)	(4.9)[b]	(8.8)[c]
QADJS	29.78	54.67	21.36	−16.65
	(20.4)[c]	(25.4)[a]	(10.9)[a]	(22.6)
GHETTO	−220.36	−737.68	−289.70	−613.90
	(340.1)	(384.2)[a]	(185.7)[b]	(520.3)[c]
PCTWT	−.38	−.59	−.24	−24.34
	(0.5)	(0.3)[a]	(0.2)[c]	(26.8)
CRIME	49	1.56	44	11
	(0.8)	(1.2)[c]	(0.2)[a]	(1.2)
SCHOLP	2.67	−12.87	−5.56	−9.43
	(57.1)	(21.5)	(9.3)	(49.3)
$\bar{R}^2$	937	891	815	897

Note: R^2 values reported are the "corrected" statistic. Numbers in parentheses are standard errors of the transformed coefficients reported in table.

[a]Coefficient significant at .05 level

[b]Coefficient significant at .10 level

[c]Coefficient has t-statistic above 1.0

run using each of the functional forms given in (8)–(11), with the appropriate u specification as in (12), (13). Results are reported only for nonwhite strata, and only for those functional forms producing the "best" results in terms of statistical significance and expected sign of coefficients (best forms are noted parenthetically). Table 2 does not show the regression coefficients directly but rather converts the coefficients to an annual dollar value for an incremental change in each $[q_i]$ component, evaluated at the mean component value of the stratum.[12] These results show that the quantitative *AGE* and *PARCL* components and qualitative *QRENV* and *QADJS* components generally wielded the largest explanatory power in most nonwhite strata. The public service variables proved of little significance. Finally, the components proxying for neighborhood racial composition (*PCTWT*) and the ghetto environment (*GHETTO*) were of particular interest, especially in light of the criticism of existing specifications in Section I above. The *PCTWT* was only sporadically statistically significant but in all cases it weakly indicated that nonwhites tended to prefer nonwhite neighborhoods to white ones.[13] *GHETTO* proved significant in two of four nonwhite strata (insufficient *GHETTO* observations were available to permit estimation for white strata) and indicated a marked aversion to this portion of the St. Louis housing market. The afore-

[12]Except for *AREA* and *PARCL* which were incremented by 100 square feet, and the quality indices which were incremented by 0.1. The Y variable used to proxy for u as in (12) and (13) was of course highly significant in every equation. The estimated coefficients for t and *NMOVE* were usually insignificant although, inexplicably, significant coefficients for these variables sometimes demonstrated opposite signs across a few strata. The specific coefficients for Y, t, and *NMOVE* are not presented in Table 2, both for brevity and because they yield no insights for the purpose at hand.

[13]This finding corroborates that of King and Mieszkowski and Gary Marx.

TABLE 3—HYPOTHETICAL BIDS FOR BENCHMARK WHITE UNITS

	White Strata					
	1	2	3	4	5	6
Actual Mean White Bid	$879	960	1033	1195	486	1198
Nonwhite Strata Hypothetical Bids						
1	$795	840	815	779	795	946
2	$1157	1219	1068	1004	1002	1323
3	$561	612	526	581	597	622
4	$1094	1096	1148	900	1066	1284
Other White Strata Hypothetical Bids						
Mean	$1010	1027	1100	1045	956	1146
Standard deviation	$91	121	135	164	112	74
Maximum	$1137	1234	1246	1254	1060	1259

mentioned negative sociopsychological elements of the ghetto environment apparently predominated here.[14] These findings indicate that studies which ignore these two factors may, indeed, suffer serious specification bias, as suggested above.

V. Tests for Discrimination

The bid-rent functions presented in the previous section will now be used to estimate how much each nonwhite stratum would hypothetically be willing to pay for typical units currently occupied by various white strata. Should any nonwhite strata be willing to bid significantly more for units presently occupied by whites (while remaining at their current u level), it would imply that some discriminatory constraint was being erected in the market to prevent housing from being allocated to the highest bidder. Remember, it is a crucial implication of bid-rent theory that in true competitive equilibrium no household will be able to outbid another for a unit and remain at the same level of welfare, regardless of the respective incomes or preferences involved in the comparison.

The initial test involved the application of the coefficients of each nonwhite stratum's best estimated bid-rent function to the mean values of their *own* current respective Y, $NMOVE$, t, and $k(t)$ and, in turn, to the mean values of each *white* stratum's $[q_i]$ bundle.[15] The comparisons between the hypothetical nonwhite bids thereby generated and actual white bids for benchmark units of various strata are presented in the upper portion of Table 3.[16] A cursory scanning of those results might suggest that nonwhite strata 2 and 4 frequently appeared willing to outbid whites. Such a conclusion would be premature, however, since no determination has yet been made as to the magnitude of bid divergencies which could be explained merely by market frictional disequilibrium. The measure of such disequilibrium bid divergencies employed here was to compare actual white bids to hypothetical bids for these benchmark units by *other white* strata, employing their own respective bid-rent functions in a manner comparable to that described for nonwhites above. The means and standard deviations of the appropriate five hypothetical white bids for each benchmark unit are presented in the lower portion of Table 3. They indicated that a substantial degree of bid disparity ($200–300) apparently must be tolerated, given the nature of housing market frictions. In light of this qualification, the observed nonwhite bids no longer appeared conclusively excessive for typical white-occupied units.

[14]The *GHETTO* result may also have been generated by such uncontrolled factors in the specification as population density, and/or non-linear interactive effects between dwelling, neighborhood, and public service qualities, crime, etc.

[15]Bids were also generated under the assumption that the bidder assumed the $k(t)$ and t of the white stratum occupying the unit, with no important alteration of results.

[16]It should be noted that bids estimated using the *CD* and *EXP* forms have an upward bias, as proven by Arthur Goldberger. This is not of great concern here since the major findings were not dependent on these particular functional forms.

TABLE 4—HYPOTHETICAL BIDS FOR BORDER BENCHMARK WHITE UNITS

	White Strata					
	1	2	3	4	5	6
Actual Mean White Bid	$653	754	904	826	746	662
Nonwhite Strata Hypothetical Bids						
1	$655	945	1016	690	823	1016
2	$1341	1636	1687	1061	1337	1705
3	$608	696	783	617	659	775
4	$877	1197	1252	712	1147	1260
Other White Strata Hypothetical Bids						
Mean	$554	872	935	626	856	923
Standard deviation	$132	184	320	222	236	183
Maximum	$854	1113	1452	1016	1247	1172

A dramatically different picture arose when hypothetical nonwhite bids were generated using mean characteristics of white units located in racially mixed tracts (35 percent < *PCTWT* < 80 percent) adjacent to the ghetto (*PCTWT* < 5 percent). Results of these tests for border areas are presented in Table 4 in a form comparable to Table 3. These areas bordering on the ghetto hold particular interest. Certainly whites would feel most "threatened" in these areas and would have the strongest motivation to discriminate in order to preserve the "ethnic purity" of their neighborhood. What's more, the aforementioned evidence of nonwhite distaste for predominantly white areas suggests their bids would be comparatively higher in mixed border regions, *ceteris paribus*.

As in the initial estimations a degree of disequilibrium intrawhite bid divergency was detected. But, unlike the previous case, nonwhite stratum 2 appeared to be significantly discriminated against in border areas, even granting a considerable degree of disequilibrium bid divergency. For every benchmark white border unit (except that occupied by white stratum 4) the nonwhite middle class, small family, mature-aged household group 2 generated hypothetical bids averaging $700 more than the existing white bids.[17] Even taking a conservative approach by claiming that discrimination was only shown by the degree to which nonwhite stratum 2 bids exceeded those of the *maximum* hypothetical white bids of strata not occupying the given unit, i.e., assuming a high disequilibrium component, the results indicated an average interracial bid divergency of $400 in border regions.

In order to establish that these results were not spurious, consideration must be given to how sensitive the bid estimations were to the choice of bid-rent functional form. Although the magnitude of the discriminatory impact in border areas appeared somewhat smaller when using other forms, the results generated by averaging hypothetical nonwhite bids over all four functional forms[18] still showed stratum 2 nonwhites willing to bid an average of $450 more than existing white bids (and $160 more than the maximum hypothetical bids of other whites). Even granting a $200–300 disequilibrium margin of error the gap remains dramatic. These average estimates should not, however, command as much confidence as those generated by the best (*CES*) functional form for stratum 2 since it was chosen on the basis of reasonableness and significance of coefficients before any bids were simulated.

The above estimates suggested, therefore, that the observed gap between nonwhite stratum 2 hypothetical bids and actual white bids for benchmark units in border areas was neither a

[17]Recall that the dependent variable in the *CES* formulation was not B, but $1/(Y - B - k(t))$, thus the standard error of the estimate varied somewhat depending on particular values of simulated B. But an average standard error of the estimate was $550, thereby suggesting this bid-gap was nontrivial.

[18]The standard deviation across these four hypothetical bids was $180

statistical artifact nor the result of normal market frictions. Furthermore, it is of importance to note additional evidence that reduces the likelihood that the gap may have been explained by other types of nondiscriminatory disequilibrium situations. One might suggest the hypothesis that if high proportions of stratum 2 households were recent in-migrants into St. Louis prior to 1967 that their unfamiliarity with the local housing market was the cause of their higher bid levels. The sampled stratum 2 households did not strongly support such an argument, however, since 34, 54, and 71 percent had lived in their current unit at least ten, five, and two years, respectively. Similarly, one might claim that the lack of white border occupants response to higher nonwhite bids was not due to discrimination but rather was due to an unwillingness to move from a long time residence and/or a lack of "market orientation." While such an argument had some force for the two older border white strata 5 and 6, it certainly did not for the other border whites. Only 11 percent of the other (0–55 years) sample border white households lived in their unit more than ten years, while 48 percent had moved in within the last two years. The majority of current white border units had been on the market immediately prior to the survey, yet were occupied by whites in spite of apparently higher stratum 2 nonwhite bids. While its exact form cannot be discerned from this study, discrimination clearly must have been the culprit here, not lack of information or market orientation or normal housing market frictions.

VI. Conclusion

The empirical study reported here has suggested that significant discriminatory constraints were operating in the St. Louis housing market in 1967. No alternative explanations, either statistical or behavioral, proved convincing. While it was impossible to ascertain what types of discriminatory behavior were involved, their primary impact was clearly to inhibit the effective bidding power of middle-aged, middle class, small nonwhite households in areas bordering the ghetto, and thereby to create a serious welfare loss for this group in the form of higher housing prices.

Various estimates placed this discriminatory constraint as equivalent to, at minimum, a \$160–\$400 (or 20–50 percent) average markup of white border benchmark unit annual rents, even allowing for substantial market frictions. While this magnitude of discrimination was larger than that suggested by the aforementioned econometric studies, it was certainly plausible insofar as these previous aggregative techniques could easily have masked greater impacts on particular nonwhite groups in particular spatial submarkets.

That the discrimination was visible only in racially mixed tracts bordering the ghetto was consistent with the findings of King and Mieszkowski, whereas its primary impact on middle class nonwhites was suggested earlier by Corienne Robinson and Chester Rapkin. Insofar as nonwhite stratum 2 demonstrated a significant aversion to heavily white neighborhoods it was clear why their attempts to secure housing concentrated (and was apparently met with resistance) in border areas. It could be contended that lower class nonwhites did not suffer so heavily from confinement in the ghetto submarket since it might have represented a more nearly optimal housing choice.[19] Middle class nonwhites, on the other hand, undoubtedly found their utility-maximizing housing bundles in short supply in the ghetto, whence the prices of those which did exist were bid up to a sufficiently high level that their hypothetical bids for comparable nonghetto border units appeared higher than those actually tendered by the white occupants. The evidence presented in this study suggests that discriminatory constraints kept these bids from being actualized, not a variety of market disequilibrium factors.

[19]Lower class nonwhites do, of course, suffer indirectly from discrimination against middle class nonwhites via the artificial fillip to ghetto demand (and prices) thereby created It can also be noted that one should not interpret the above markups as the amount stratum 2 rents would fall in the absence of discrimination since that would make strong presumptions about the long-run general equilibrium nature of the housing market which have not been proven

REFERENCES

William Alonso, *Location and Land Use,* 4th ed., Cambridge 1970.

M. Bailey, "Effects of Race and of Other Demographic Factors on the Value of Single Family Homes," *Land Econ.,* Aug. 1959, *42,* 215–20.

David Birch et al., *America's Housing Needs: 1970–1980,* Cambridge 1973.

Kenneth Clark, *Dark Ghetto,* New York 1965.

C. Daniels, "The Influence of Racial Segregation on Housing Prices," *J. Urban Econ.,* Apr. 1975, *2,* 105–22.

Beverly Duncan and Philip Hauser, *Housing a Metropolis: Chicago,* Glencoe 1960.

G. Galster, "A Bid-Rent Analysis of Housing Market Discrimination," unpublished doctoral dissertation, M.I.T. 1974.

———, "Prejudice vs. Preference: What Do We Really Know About Housing Market Discrimination?" *Reg. Sci. Perspectives,* 1976, *6,* 17–27.

Herbert Gans, *The Urban Villagers,* New York 1962.

A. Goldberger, "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica,* Oct. 1968, *36,* 464–72.

Ulf Hannerz, *Soulside: Inquiries Into Ghetto Culture and Community,* New York 1969.

B. Harris, J. Nathanson, and L. Rosenberg, "Research on an Equilibrium Model of Metropolitan Housing and Locational Change: Interim Report," Planning Sci. Group/Inst. Environ. Stud., Univ. Pennsylvania, Mar. 1966.

LeRoi Jones, *Blues People,* New York 1963.

J. Kain and J. Quigley, "Measuring the Value of Housing Quality," *J. Amer. Statist. Assn.,* June 1970, *62,* 532–48.

——— **and** ———, "Housing Market Discrimination, Homeownership, and Savings Behavior," *Amer. Econ. Rev.,* June 1972, *62,* 263–77.

——— **and** ———, *Housing Markets and Racial Discrimination: A Microeconomic Analysis,* New York 1974.

A. T. King and P. Mieszkowski, "Racial Discrimination, Segregation, and the Price of Housing," *J. Polit. Econ.,* May/June 1973, *81,* 590–606.

J. Lansing and L. Kish, "Family Life Cycle as an Independent Variable," *Amer. Sociological Rev.,* Oct. 1957, *22,* 512–19.

V. Lapham, "Do Blacks Pay More for Housing?," *J. Polit. Econ.,* Nov./Dec. 1971, *79,* 1244–57.

Gary Marx, *Protest and Prejudice,* New York 1967.

Daniel Moynihan and Nathan Glazer, *Beyond the Melting Pot,* Cambridge 1963.

L. Moses, "Towards a Theory of Intra-Urban Wage Differentials and their Influence on Travel Patterns," *Papers Proc. Reg. Sci. Assn.,* 1962, *9,* 56–63.

E. Olsen, "Do the Poor or the Black Pay More for Housing?" in George von Furstenberg et al., eds., *Patterns of Racial Discrimination,* Vol. 1, Lexington 1974, 205–11.

Lee Rainwater, *Behind Ghetto Walls,* Chicago 1970.

Chester Rapkin, "Price Discrimination Against Negroes in the Rental Housing Market," in *Essays in Urban Land Economics,* Los Angeles 1966.

——— **and William Grigsby,** *The Demand for Housing in Racially Mixed Areas,* Berkeley 1960.

R. Ridker and J. Henning, "The Determinants of Residential Property Values with Special Reference to Air Pollution," *Rev. Econ. Statist.,* May 1967, *49,* 246–57.

C. Robinson, "The Relationship Between Condition of Dwelling and Rentals, By Race," *J. Land Publ. Util. Econ.,* Aug. 1946, *22,* 296–302.

M. Straszheim, "Estimation of the Demand for Urban Housing Services from Household Interview Data," *Rev. Econ. Statist.,* Feb. 1973, *55,* 1–6.

———, "Housing Market Discrimination and Black Housing Consumption," *Quart. J. Econ.,* Feb. 1974, *88,* 19–43.

W. Wheaton, "Income and Urban Location," unpublished doctoral dissertation, Univ. Pennsylvania 1972.

Intertemporal Utility Maximization and the Timing of Transactions

By PETER HOWITT*

This paper addresses the problem of explaining a household's choice of consumption and purchasing plans on the basis of a model of intertemporal utility maximization. Until recently the intertemporal choice models that have been developed by economists have simply not distinguished between purchasing, a market activity, and consumption, a nonmarket activity.[1] The importance of this distinction, and of incorporating it into a model of intertemporal choice, can be seen from the point of view of three separate areas of current research.

First, recent work on the microfoundations of monetary theory[2] has shown the importance of transaction costs in explaining the role of money in economic activity. The absence of these costs from standard general equilibrium theory makes it difficult to account for the special characteristics, and even the existence, of money within that framework. This research has underlined the need to develop a theory of transactions on the same level of sophistication as our theories of production and consumption. The present problem may be viewed as one part of the larger problem of developing such a theory of transactions.

Second, in the area of short-run aggregate analysis, purchasing decisions are of more interest than consumption decisions because they are more closely related to the level of aggregate demand. Recent empirical investigations by Michael Darby have supported Milton Friedman's conjecture that the aggregate rate of purchase of consumer durables can undergo large fluctuations even when the aggregate rate of consumption is relatively constant. It would clearly further our understanding of short-run fluctuations in aggregate demand if both of these rates could be explained on the basis of intertemporal choice.

Third, the area most closely related to the present problem is the inventory theory of the demand for money. This theory has been extended in recent years by several authors into a generalized theory of the size and timing of all sorts of transactions, including wage payments, commodity purchases and sales, and various financial transactions.[3] While this research has made considerable progress in developing a separate theory of transactions, it has not been related explicitly to models of intertemporal choice.[4] In the absence of such a model, work in this area has been exclusively concerned with stationary-state phenomena. The familiar square root formulae are only valid in a situation where all planned consumption and production flows and planned transaction sizes and frequencies are constant.[5] The present approach may be regarded as a generalization of inventory theory to cover nonstationary situations in which these

*University of Western Ontario. I am indebted to Robert Clower who provided the original stimulus for this paper, and to Joel Fried, with whom I have had many fruitful discussions on the subject. Helpful suggestions have been provided by Michael Parkin and an anonymous referee. I have benefitted from having seen the results of some preliminary work done in 1972 on a similar topic by D. W. Bushaw and others at Washington State University. The research was financed by a grant from the University Research Council of the University of Western Ontario.

[1]See, for example, the papers by Miguel Sidrauski and Hirofumi Uzawa.

[2]See, for example, Karl Brunner and Allan Meltzer, Robert Clower (1971), and Joseph Ostroy.

[3]See Robert Barro and Anthony Santomero, Clower (1970), and Edgar Feige and Michael Parkin.

[4]See, however, the unpublished work by Constantino Lluch and Henri Lorie that has begun to explore explicit intertemporal formulations of the theory.

[5]George Hadley and T. M. Whitin, pp. 22–24, 29–40, 336–45, present a good account of the stationary nature of the square root formula as applied to problems in management science.

planned magnitudes are changing over time.

There is a major analytical difficulty that must be overcome in order to build such a model. An essential feature of inventory theory is the presence of a set-up cost of transacting—a cost incurred at each transaction date but which is independent of the size of the transaction. This cost prevents an agent from transacting continuously through time, for to do so would be to make an infinite number of transactions, and hence incur an infinitely large cost, over a finite interval of time. Furthermore, everyday observation suggests that the time path of the typical household's purchases is more lumpy than the time path of its consumption. This is attributable partly to the indivisibility of many durable goods but partly also to the indivisibility of transaction costs. Thus it is natural to regard the household as consuming continuously through time but purchasing at discrete points in time. But this implies that the household's choice problem involves a potentially intractable mixture of discrete and continuous time analysis. The stationary approach of inventory-theory avoids this difficulty by defining the objects of choice as single valued rates of consumption and production and a single valued transaction frequency for each good. In an explicit intertemporal model the problem presents itself in the form of what control theorists call "bang-bang" analysis—the analysis of a dichotomous choice between either doing something in some finite amount or doing nothing at all, with choices in between being ruled out. In the present context the choice must be made at each date either to transact at least enough to cover the transaction cost or not to transact at all.

The specific objective of the present paper is to suggest an approach to this general problem by analyzing the special case in which the household deals in only two commodities. The plan of the paper is as follows. Section I outlines the general nature of the household's decision problem and contrasts it to the analogous stationary problem of inventory theory. Section II presents the formal analysis of the problem and shows how by solving the problem in stages one can reduce the mixed continuous-discrete time problem to a relatively tractable problem of discrete dynamic programing. Section III characterizes the solutions to the problem by analyzing the existence, uniqueness, and stability of stationary solutions and direction of response of the short-run solutions to parameter changes. Section IV compares the present approach to that of inventory theory by demonstrating how the stationary solution in the former may be approximated by the square root formula of the latter. Section V concludes by discussing extensions and applications.

I. The Household Decision Problem

Consider the situation of a household that deals in just two commodities, an income good Y and a consumption good C. At any date t along a continuous time scale the household possesses inventories of these two commodities, the quantities of which are denoted by $Y(t)$ and $C(t)$, respectively. The income good is produced continuously at the constant flow rate of y units per unit-time, and the consumption good is consumed continuously at the flow rate of $c(t)$ units per unit-time. There is no production of the consumption good or consumption of the income good. Thus, over any time interval during which no market transaction occurs the rates of change of the stocks $Y(t)$ and $C(t)$ will be y and $-c(t)$, respectively. In line with the discussion above it is assumed that transactions occur discontinuously at discrete points in time. At any transaction date, the household may sell an amount of the income good no greater than the existing inventory at a constant rate of exchange (assumed for notational simplicity to be unity) for the consumption good. At each transaction date the household must pay to a broker a transaction cost in the form of α units of the income good, where α is a fixed amount, independent of the size of the transaction. Thus at each transaction date $Y(t)$ will decrease discontinuously by the amount of the sale Q, and $C(t)$ will increase by the amount of the sale minus the transaction cost $Q - \alpha$. The household's problem is to choose its time path of consumption,

and the size and timing of each transaction, subject to the constraints indicated above as well as the obvious constraint that $C(t)$ cannot become negative; that is, once $C(t)$ becomes zero no consumption may occur until the next transaction date.

There are several interpretations that can be placed upon this situation. The household may be imagined as being in a barter economy where neither of the goods has the property of medium of exchange, or in a monetary economy where either of the two goods is money. If the income good is money, the household is being paid continuously for a constant flow of services. If the consumption good is money (the assumption originally made by William Baumol and James Tobin) the household is making continuous (costless) purchases with the flow rate of purchase equal to the flow rate of consumption.

In choosing the timing of its transactions the household must strike a balance between transaction costs and waiting costs. The total transaction cost within any interval of time will be α times the number of transactions within that interval. Thus the greater the frequency of transactions the larger the average rate of transaction cost per unit-time. On the other hand the amount of the consumption-good inventory required at the beginning of any transaction interval (the interval between two consecutive transaction dates) to accommodate a given average rate of consumption $\bar{c}$ throughout that interval is equal to $\bar{c}$ times the length of the interval. Since waiting costs must be incurred in order to begin any interval with a larger inventory we may say that the higher the frequency of transactions the shorter will be the average transaction interval and hence the lower the average rate of waiting costs per unit-time. Thus the household will choose to increase the frequency of transactions up to the point at which the marginal reduction in waiting costs is just offset by the marginal increase in transaction costs.

The stationary approach of inventory theory focuses exclusively on this particular tradeoff by assuming that consumption is predetermined at a rate just large enough to maintain constant stocks; in other words, if the necessary deduction for payment of the transaction costs is ignored, $c(t) = y$ for all t. The household's problem in this context is to choose a constant value of λ, the length of its transaction intervals (the frequency of transactions is $1/\lambda$), so as to minimize the sum of average transaction costs per unit-time C_T, and average waiting costs per unit-time C_W. The value of C_T will be the cost per transaction times the frequency of transactions α/λ. The value of C_W may be expressed as the average value of inventories times the constant rate of time discount ρ. Just before a transaction occurs the household's inventories will consist entirely of the amount $y\lambda$ of the income good, the amount that has been produced over the transaction interval. (This assumes, of course, that no precautionary balances of either good are carried over between transaction intervals.) But since the household is in a stationary state the total value of its inventories will be constant over time, at the amount $y\lambda$. Thus the value of C_W will be $\rho y\lambda$. The optimal value of λ will be that which minimizes $\alpha/\lambda + \rho y\lambda$:

$$\lambda = \sqrt{\alpha/\rho y} \tag{1}$$

In contrast, an explicit intertemporal approach postulates initial stock holdings given by the past. For example, assume that the household begins with a zero holding of the income good and some positive quantity C_0 of the consumption good. If it consumes throughout the present transaction interval at the average rate of $yC_0/(C_0 + \alpha)$ then when it has just exhausted its initial stock of the consumption good it will have accumulated enough of the income good to begin the next transaction interval in exactly the same position as it began the current one. This average rate of consumption is the permanently maintainable rate given the initial holding C_0. If C_0 is very small the permanently maintainable rate will be very small because of the necessity for very frequent transactions which implies a high average transaction cost per unit of time. The household may, in this case, decide to consume at a lower than permanently maintain-

able average rate during the current transaction interval. If this happens then by the time the initial holding is exhausted the household will have accumulated enough of the income good to begin the next transaction interval with more than C_0. On the other hand if the initial stock is very large the household may decide that the high permanently maintainable rate permitted by this stock is not enough to discourage it from consuming at an even higher rate, in which case it will begin the next interval with a smaller holding.

From this point of view the household's transaction decisions will be governed by an intertemporal tradeoff. It can acquire current consumption at the expense of the future by shortening the transaction interval, or give up current consumption to gain more in the future by lengthening it.

II. Formalizing the Problem

The present section presents a formalization of this tradeoff. Suppose the household's preferences are given by the additive utility functional:

$$U = \int_0^{\infty} u(c(t))e^{-\rho t}dt \tag{2}$$

with the twice continuously differentiable function u satisfying:

$$u' > 0,\ u'' < 0,\ \lim_{c \to 0} u'(c) = \infty \tag{3}$$

The choice-objects are the consumption plan $c = \{c(t): t \geqq 0\}$, the set of transaction dates T, and the corresponding set of transaction quantities $Q = \{Q(t): t \in T\}$. Let C_0 and Y_0 denote the initial stocks. The time paths of stocks must obey:

$$C(t) = C_0 + \sum_{\{\tau \leqq t,\ \tau \in T\}} [Q(\tau) - \alpha] - \int_0^t c(t)dt \geq 0 \tag{4}$$

$$Y(t) = Y_0 - \sum_{\{\tau \leqq t,\ \tau \in T\}} Q(\tau) + \int_0^t y dt \geq 0 \tag{5}$$

for all $t \geqq 0$. The household will attempt to maximize U subject to (4) and (5), given C_0 and Y_0.

This mixed discrete-continuous time problem can be given a tractable formulation in three steps. The first step is to assume that $Y_0 = 0$; i.e., date 0 is a transaction date. This adds notational simplicity with no loss in generality, for the case of $Y_0 > 0$ can be treated by going back to the last transaction date before date 0. The second step is to note that the solution must always satisfy the following three properties. First, the set T must be countable; i.e., it consists of a sequence $\{t_1, t_2, \ldots\}$, for otherwise there would be an infinity of transactions over some finite time interval and (4) would necessarily be violated during that interval. Second, at each transaction date the household will sell all of its accumulated inventory of the income-good; i.e., $Y(t_n) = 0,\ n = 1, \ldots, \infty$. This obviously holds because failure to sell all of the income good will not save holding costs, as the goods are equally costly to store, whereas it will require the next transaction to occur sooner than otherwise.[6] This property implies that $Q(t_n) = y(t_n - t_{n-1}),\ n = 1, \ldots, \infty$ (where $t_0 = 0$). Third, because there are no differential holding costs and no uncertainty to generate a precautionary inventory demand, no transaction will occur unless the stock of consumption good has been exhausted.[7] These properties imply that the consumption good inventory after each transaction date will be $C_n \equiv C(t_n) = y(t_n - t_{n-1}) - \alpha$.

The third step is to note that over any transaction interval $[t_n, t_{n+1}]$ during which the

[6]Formally, if any plan $\{c^*, T^*, Q^*\}$ did not satisfy this property, utility would not be reduced by replacing it with the plan $\{c^*, T^*, Q^{**}\}$, where

$$\sum_{i=1}^{n} Q^{**}(t^*_i) = \int_0^{t^*_n} y dt,\ n = 1, \quad , \infty$$

[7]In this case any optimal plan $\{c^*, T^*, Q^*\}$ could be replaced by the plan $\{c^*, T^{**}, Q^{**}\}$, where

$$\int_0^{t^{**}_1} y dt = Q^{**}(t^{**}_1),\quad \int_0^{t^{**}_1} c^*(t)dt = C_0$$

and

$$\int_{t^{**}_{n-1}}^{t^{**}_n} y dt = \int_{t^{**}_n}^{t^{**}_{n+1}} c^*(t)dt + \alpha = Q^{**}(t^{**}_n),\ n = 2, \ldots, \infty$$

amount C_n is consumed, the consumption plan must solve the problem:

$$(6) \qquad \text{Max} \int_0^{\lambda_{n+1}} u(c(t_n + t))e^{-\rho t}dt$$

subject to $\int_0^{\lambda_{n+1}} c(t_n + t)dt = C_n$

where $\lambda_{n+1} \equiv t_{n+1} - t_n$ is the length of the interval. The solution to (6) is determined by the constraint and the Euler condition:

$$(7) \qquad \frac{d}{dt} u'(c)e^{-\rho t} = 0$$

which states that the discounted marginal utility of consumption must be equalized throughout the interval. This condition requires the rate of consumption to fall steadily throughout the interval, even in a steady state with λ constant. The optimized value of the objective function in (6) may be written as the function $J(\lambda_{n+1}, C_n)$. This function expresses the maximum attainable utility from consuming the amount C_n over an interval of length λ_{n+1}, where utility has been discounted to the beginning of the interval.

These three steps permit the decision problem to be formulated as:[8]

$$(8) \qquad \text{Max}\, J(\lambda_1, C_0) + \sum_{n=1}^{\infty} J(\lambda_{n+1}, y\lambda_n - \alpha) \exp\left\{-\rho \sum_{i=1}^{n} \lambda_i\right\}$$

where the objective function consists of the sum of utilities attainable within each interval, and the choice-objects are the interval lengths themselves. Let $V(C_0)$ represent the optimized value of the objective function in (8); i.e., the utility attainable over an infinite horizon beginning with the amount C_0. This can be decomposed into two parts: the utility attainable over the first interval and the utility attainable from t_1 on. The latter is just the utility attainable over an infinite horizon beginning with the amount $y\lambda_1 - \alpha$, discounted to date zero. Thus we may rewrite (8) as[9]

$$(9) \qquad V(C_0) = \text{Max}\,\{J(\lambda_1, C_0) + e^{-\rho\lambda_1} V(y\lambda_1 - \alpha)\}$$

The first-order condition states:[10]

$$(10) \qquad J_1(\lambda, C) - \rho e^{-\rho\lambda} V(y\lambda - \alpha) + ye^{-\rho\lambda} V'(y\lambda - \alpha) = 0$$

Increasing λ has three effects, represented by the three terms of (10). The first is the gain during the first interval of making it longer;[11] the second is the subjective interest cost of deferring all future consumption, and the third is the gain from beginning the next interval with a larger stock.

The formal structure of this discrete dynamic programing problem is different from the sort familiar to economists in that the length of the period is a choice-object. There are three instances in economics where similar intertemporal problems have been discussed. The first is the Wicksellian capital model[12] in which the choice-object is the durability of capital goods. The present model is formally different from the Wicksellian model in that the choice of λ will generally be dependent upon past choices as embodied in C. The second is J. S. Flemming's article on consumer durable purchases. It was limited to the case of constant elasticity of marginal utility, and assumed that the rate of consumption during any interval was automatically determined by the size of the last purchase. The third is the paper on optimal growth by Avinash Dixit, James Mirrlees, and Nicholas Stern, in which the choice-object was the interval between successive undertakings of discrete investment projects in the presence of increasing returns. It assumed no time-discounting.

III. Characterizing the Solutions

A

The first-order condition (10) can be solved for the first interval length as a function $\lambda^*(C)$

[8]The assumption (3) allows us to ignore the nonnegativity constraints $y\lambda_n - \alpha \geqq 0$ in (8).

[9]This argument is just an application of Bellman's principle of optimality, p. 83.

[10]For notational simplicity the subscripts on λ and C are dropped whenever no ambiguity results.

[11]This may be positive or negative. See Appendix A.

[12]For a modern restatement of the Wicksellian model, see David Cass.

of the beginning of period stock. It can be shown that $d\lambda^*/dC > 0$.[13] Therefore the household's behavior over time will be governed by the difference equation

$$\lambda_n = \lambda^*(y\lambda_{n-1} - \alpha), \qquad n = 1, \ldots, \infty \tag{11}$$

where the initial position is determined by C_0

$$\lambda_0 = (C_0 + \alpha)/y \tag{12}$$

A stationary solution is defined as a rest point of this system, i.e., a point $\hat{\lambda}$ such that $\hat{\lambda} = \lambda^*(y\hat{\lambda} - \alpha)$.

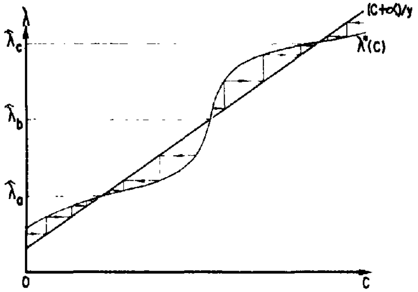


Figure 1

Appendix B demonstrates that a stationary solution exists, and that λ_n will converge upon a stationary solution from any initial position. The household's behavior can be described by Figure 1, where $(C + \alpha)/y$ is the "permanently maintainable" value of λ. Given the current choice λ^*, the value of C next period will be $y\lambda^* - \alpha$, which is just the value of C that would make λ^* permanently maintainable. In the case of multiple stationary solutions as in Figure 1, stable and unstable solutions alternate, but every motion will converge on some stationary solution and the convergence will be monotonic in λ and C.

This dynamic behavior may be interpreted in the terms of Section I. Given any value of C the average rate of consumption during the first interval will be $\bar{c} = C/\lambda^*(C)$. Whenever $\lambda^*(C) < (C + \alpha)/y$, then $\bar{c} > yC/(C + \alpha)$; that is, average consumption will be more than the permanently maintainable rate and inventories will be falling. Whenever $\lambda^*(C) > (C + \alpha)/y$, average consumption will be at less than the permanently maintainable rate and inventories will be rising.

B

The optimal choice of the first interval length and of the average rate of consumption during the first interval may be expressed as functions of the initial inventory, the cost per transaction, and the level of income: $\lambda(C, \alpha, y)$ and $\bar{c}(C, \alpha, y)$. The second of these may be interpreted as a short-run consumption function. The rest of this section investigates the qualitative manner in which the short-run solutions respond to changes in C, α, and y. The analysis will be restricted to a neighborhood of a stable stationary solution (*SSS*). The nature of the responses in such a neighborhood can be inferred from the signs of the partial derivatives evaluated at an *SSS*. In the case of α and y the method of investigation is to consider the second derivatives of an indirect utility function, making use of the envelope theorem. This method is of some interest in itself.

First, it is obvious from Figure 1 that at an *SSS*:

$$0 < \frac{\partial \lambda}{\partial C} < \frac{1}{y} \tag{13}$$

From the definition of $\bar{c}$, at an *SSS*:

$$\frac{\partial \bar{c}}{\partial C} = \frac{1}{\lambda^2}\left[\lambda - C\frac{\partial \lambda}{\partial C}\right] = \frac{1}{\lambda^2}\left[\lambda\left(1 - y\frac{\partial \lambda}{\partial C}\right) + \alpha\frac{\partial \lambda}{\partial C}\right] > 0 \tag{14}$$

To derive the responses to α, denote by $W(C, \alpha)$ the value of $V(C)$ given α. Then (9) may be rewritten as:

$$W(C, \alpha) = \text{Max } J(\lambda, C) + e^{-\rho\lambda} W(y\lambda - \alpha, \alpha) \tag{15}$$

The envelope theorem implies that:

[13] From (10), $H(d\lambda^*/dC) + J_{21} = 0$, where $H = J_{11} + \rho^2 e^{-\rho\lambda} V - 2\rho y e^{-\rho\lambda} V' + y^2 e^{-\rho\lambda} V''$. The necessary second-order condition for a regular maximum in (9) is $H < 0$, and Appendix A demonstrates that $J_{21} > 0$.

(16) $$W_1(C, \alpha) = J_2(\lambda, C)$$

(17) $$W_2(C, \alpha) = e^{-\rho\lambda}[-W_1(y\lambda - \alpha, \alpha) + W_2(y\lambda - \alpha, \alpha)]$$

Differentiating (16) with respect to α produces:

(18) $$W_{12}(C, \alpha) = J_{12}(\lambda, C)\frac{\partial\lambda}{\partial\alpha}$$

Appendix A shows that $J_{12}(\lambda, C) > 0$. Therefore $\partial\lambda/\partial\alpha$ has the same sign as $W_{12}(C, \alpha)$. Assume that the second derivatives of W are continuous. Differentiating (17) with respect to C, and using Young's theorem, produces:

(19) $$W_{12}(C, \alpha) = \frac{\partial\lambda}{\partial C}[-\rho W_2(C, \alpha) + e^{-\rho\lambda}y(W_{12}(y\lambda - \alpha, \alpha) - W_{11}(y\lambda - \alpha, \alpha))]$$

At an SSS, $y\lambda - \alpha = C$. Therefore, from (16), (17), and (19):

(20) $$W_{12}(C, \alpha) = \left(1 - e^{-\rho\lambda}y\frac{\partial\lambda}{\partial C}\right)^{-1}\frac{\partial\lambda}{\partial C}e^{-\rho\lambda}[\rho(1 - e^{-\rho\lambda})^{-1}J_2(\lambda, C) - yW_{11}(C, \alpha)]$$

Appendix A shows that $J_2(\lambda, C) > 0$ and that $W_{11}(C, \alpha) < 0$ at an SSS. Therefore, (13) and (20) imply $W_{12}(C, \alpha) > 0$, which implies $\partial\lambda/\partial\alpha > 0$. From the definition of $\bar{c}$, $\partial\bar{c}/\partial\alpha < 0$. By exactly analogous reasoning, $\partial\lambda/\partial y < 0$, $\partial\bar{c}/\partial y > 0$.

IV. The Square Root Formula

A square root formula like (1) that is derived from inventory theory is usually regarded as an approximation to the stationary solution of an underlying intertemporal model like the present one. The present section investigates the question of the validity of this approximation in the present model. There are two distinct questions involved. One is the question of whether the stationary solution of the present model (assuming that it is unique) may be approximated closely by (1), and the second is whether the nonstationary solutions closely approximate the stationary solution. The second question is that of the speed of convergence of the system (11). The greater the speed of convergence the more likely it will be that a household's observed choices will be close to a stationary solution, and hence the less will be the likely error involved in approximating the nonstationary solution by the stationary solution.

The answer to the first question depends upon the approximate validity of two simplifying assumptions of inventory theory. One is that variations in consumption levels within each transaction interval may be ignored. The other is that the compounding of interest within each transaction interval may be ignored. Assume that the household is in a stationary state. The first-order condition (10) may be rewritten as:[14]

(21) $$u(x) - xu'(x) - \rho V(C) + ye^{-\rho\lambda}u'(x) = 0$$

where x is the rate of consumption at the very end of the interval. If the first of these simplifying assumptions is made, then $x = \bar{c} = y - \alpha/\lambda$, the constant rate of consumption, and $\rho V(C) = \rho\int_0^\infty u(\bar{c})e^{-\rho t}dt = u(\bar{c})$. Thus (21) reduces to

(22) $$y(e^{-\rho\lambda} - 1) + \alpha/\lambda = 0$$

The second simplifying assumption allows $(e^{-\rho\lambda} - 1)$ to be replaced by $-\rho\lambda$, and (1) follows directly from making this replacement in (22). Note that this last step is equivalent to approximating (22) by a first-order Maclaurin series in ρ.

Ignoring the variation in consumption levels within each interval can be justified if the elasticity of marginal utility, $-cu''(c)/u'(c)$, is large enough Given any values of ρ and λ the drop in consumption throughout the interval can be made arbitrarily small by assuming that this elasticity is arbitrarily large. Ignoring the compounding of interest can be justified if $\rho\lambda$ is small enough. If the stationary solution is closely approximated by (1) then $\rho\lambda$ will be small when ρ is small. Whether $\rho\lambda$ does indeed become small when ρ is small is an open question. Even if it were answered in the affirmative there remain the questions of how large an elasticity or how low a discount rate would be required to produce a good approximation. This can probably best be answered by performing numerical calculations with specific utility functions.

Without actually performing these calculations the following example suggests that the closeness of the approximation may be sensitive

[14]See Appendix A.

to the elasticity of marginal utility. Consider the limiting case where this elasticity is zero. Assume, with no loss in generality, that $u(c) = c$. In this case there is, strictly speaking, no solution to the problem (6). The Euler equation (7) requires the marginal utility to rise throughout the interval, but in this case the marginal utility is a constant. However it is clear on economic grounds that the household will concentrate all of the consumption during any interval on the first date of the interval, for that will always be the date of the greatest discounted marginal utility. Because of this lumpiness in consumption the zero-elasticity case is probably not very meaningful in itself, for it is inconsistent with the everyday observations referred to above in the introduction. But it does serve as an extreme example to suggest what might happen when the elasticity is not large enough for the first simplifying assumption to be valid.

A simple limiting process establishes the result:[15]

$$J(\lambda, C) = C \tag{23}$$

The first-order condition (10) can be written as:

$$-\rho V(y\lambda - \alpha) + y = 0 \tag{24}$$

The solution to (24) is independent of C; i.e., the household in this case goes immediately from any initial position to the stationary solution. It follows that:

$$V(y\lambda - \alpha) = \sum_{n=0}^{\infty} e^{-\rho n\lambda} J(\lambda, y\lambda - \alpha) = \frac{y\lambda - \alpha}{1 - e^{-\rho\lambda}} \tag{25}$$

From (24) and (25),

$$-\rho(y\lambda - \alpha) + y(1 - e^{-\rho\lambda}) = 0 \tag{26}$$

Making the second simplifying assumption as before produces in this case the nonsense result: $\rho\alpha = 0$. However, taking the Maclaurin series to the second-order results in the equation

$$\rho\alpha - \rho^2 y\lambda^2/2 = 0 \tag{27}$$

[15]Suppose that the amount C was consumed at a constant rate (C/k) for an interval of duration k within the overall transaction interval. Then the utility within that interval would be $\int_0^k (C/k)e^{-\rho t}dt = C(1 - e^{-\rho k})/\rho k$. The result (23) is the limiting value of utility as k approaches zero

and the formula

$$\lambda = \sqrt{2\alpha/\rho y} \tag{28}$$

This square root formula, while not the same as (1), is exactly the formula that is arrived at by following an argument analogous to the one by which (1) was derived in Section I, noting that in this case the average inventory holdings will only be $y\lambda/2$ because no consumption-good inventories are ever held.

Three general suggestions may be drawn from this extreme example. First, the closeness of the approximation (1) appears to be sensitive to the elasticity of marginal utility. The example suggests that a low elasticity will make the stationary value of λ larger than (1) because of the implied reduction in average inventory holdings. Second, the speed of convergence in this case is infinite, which suggests that the closeness of the approximation (1) to the stationary solution is inversely related to the speed of convergence. Third, the fact that a square root formula is derivable even in this extreme case suggests that one important result of inventory theory, that the income elasticity of inventory holdings should be one-half, is valid under quite general assumptions regarding utility as a property of the stationary solution for small enough values of ρ.

V. Extensions and Applications

The approach and techniques of the present paper are applicable to a broader class of models than the present one. For example one could assume that the household's income is received in the form of returns from holdings of a perpetuity that compounds continuously at the rate r, and that there is a set-up cost of selling the perpetuity. The household's problem starting at some transaction date will be to choose a sequence of interval lengths, $\{\lambda_1, \ldots, \lambda_n, \ldots\}$ and a sequence of encashment sizes, $\{M_1, \ldots, M_n, \ldots\}$ so as to:

$$\text{Max } J(\lambda_1, M_0) + \sum_{n=1}^{\infty} J(\lambda_{n+1}, M_n) \, exp\left\{-\rho \sum_{i=1}^{n} \lambda_i\right\} \tag{29}$$

$$\text{Subject to } A_n = (A_{n-1} - M_{n-1} - \alpha)e^{r\lambda}$$

$$A_0 \text{ given}$$

where A_n is the holding of the perpetuity just

before t_n. This problem is the same as the one studied by Baumol, except that we are allowing for an infinite horizon, the compounding of interest, variations in consumption within the transaction interval, and nonstationarity of the interval length. The above techniques can be applied to this problem to yield the following results. Stationarity requires that $\rho = r$. If this holds, the household goes immediately to a stationary solution. Otherwise it accumulates or decumulates indefinitely. The square root formula (1) is an approximation to the stationary solution in the same sense as in Section IV above.

Another possible extension is to the case of more than two goods. This can be done, but the solutions are difficult to characterize without the standard a priori restriction that relative trading frequencies be integer valued. While some progress has been made in relaxing this restriction in the approach of inventory theory[16] it remains to be seen whether similar progress can be made in the intertemporal approach.

One of the problems to which this approach might be applied is that of providing a more precise theoretical explanation of the differential impact of transitory income on consumption and spending referred to in the introduction. This application would require replacing the constant income stream in the present model with a variable one. Another problem is that of determining the differential wealth effects of changes in different asset holdings. It is well known that in the presence of lumpy transaction costs optimal spending decisions depend upon the entire configuration of asset holdings, not just total wealth. The present approach, by focusing attention on transaction dates, has only considered one kind of asset—consumption good inventories; but it would be a simple extension to assume that the initial date is generally not a transaction date, in which case the decisions will depend upon both C_0 and Y_0. Finally, this approach could be applied to the problem of aggregating over households. Several recent authors have suggested that aggregate demands for assets ought to respond smoothly to discrete parameter changes because of the presence of transaction costs.[17] Transaction costs, unlike the adjustment costs involved in the theory of investment demand, are not convex, and do not lead to gradual adjustment for an individual trader, but smooth aggregate responses may be derivable nevertheless from the statistical properties of aggregation.[18] The present approach could be used to study such aggregate responses by assuming, for example, that households are identical except that at any time they are at different points within the transaction interval, with transaction dates that are smoothly distributed over households.

[16]See Clower and Howitt.

[17]See, for example, Douglas Purvis.

[18]This has been done in a simple stationary case by Robert Barro

APPENDIX

A

This Appendix proves four statements made in the text:

(A1) $J_{12}(\lambda, C) > 0$

(A2) $J_2(\lambda, C) > 0$

(A3) $W_{11}(C, \alpha) < 0$ at an SSS

(A4) Equation (10) may be rewritten as: $u(x) - xu'(x) - \rho V(C) + ye^{-\rho}u'(x) = 0$ at a stationary solution, where x is the rate of consumption at the end of the interval.

Let $\{z(t; \lambda, C); t \in [0, \lambda]\}$ be the solution to (6). Then $x = z(\lambda; \lambda, C)$. Standard techniques of the calculus of variations produce the results: $J_1(\lambda, C) = [u(x) - xu'(x)]e^{-\rho\lambda}$ and $J_2(\lambda, C) = u'(x)e^{-\rho\lambda}$ (see Hadley and M. C. Kemp, pp. 117–20). Statement (A2) follows immediately; (A4) follows from these results and the envelope result; $V'(C) = J_2(\lambda, C)$. Differentiating J_1 and J_2 with respect to C produces: $J_{12} = -xu''(x)e^{-\rho\lambda}\,\partial z/\partial C$ and $J_{22} = u''(x)e^{-\rho\lambda}(\partial z/\partial C)$ It is easily shown that $\partial z/\partial C > 0$, from which (A1) follows From the envelope result (16), $W_{11} = J_{12}(\partial\lambda/\partial C) + J_{22} = u''(x)e^{-\rho\lambda}(\partial z/\partial C)(1 - x(\partial\lambda/\partial C))$. At an SSS, (13) holds and $x < y$, from which (A3) follows

B

This Appendix proves that a stationary solu-

tion exists, and that the system (11) converges from any initial position. From Figure 1 it is obviously sufficient to prove that there is no $\bar{C}$ such that $\lambda^*(C) > (C + \alpha)y$ for all $C \geqq \bar{C}$ (this admits the possibility that the only stationary solution is $\lambda = \alpha/y$). Suppose that the function $u(c)$ is bounded, with $u^* = \lim_{c\to\infty} u(c)$ (the proof in the case of unbounded utility follows analogously). Suppose there is such a $\bar{C}$. Then for initial values $C \geqq \bar{C}$, the system (11) diverges, with $\lambda_n \rightarrow \infty$. It follows that (i) $\rho V(y\lambda_n - \alpha) \rightarrow u^*$, and (ii) $V'(y\lambda_n - \alpha) \rightarrow 0$. Since λ_n increases monotonically it also follows that $x < y$, so that (iii) $u(x) < u(y) < u^*$. As in (A4) above we may rewrite the equation (10) as $u(x) - xu'(x) - \rho V(y\lambda_n - \alpha) + yV'(y\lambda_n - \alpha) = 0$. But, from (i) – (iii), this expression must eventually become negative. The proof follows from this contradiction.

REFERENCES

R. J. Barro, "Integral Constraints and Aggregation in an Inventory Model of Money Demand," *J. Finance,* Mar. 1976, *31,* 77–87.

——— **and A. M. Santomero,** "Output and Employment in a Macro Model with Discrete Transaction Costs," *J Monet Econ.,* July 1976, *2,* 297–310.

W. J. Baumol, "The Transactions Demand For Cash: An Inventory Theoretic Approach," *Quart. J. Econ.,* Nov. 1952, *66,* 545–56.

Richard Bellman, *Dynamic Programming,* Princeton 1957.

K. Brunner and A. H. Meltzer, "The Uses of Money: Money in the Theory of an Exchange Economy," *Amer. Econ. Rev.,* Dec. 1971, *61,* 784–805.

D. Cass, "On the Wicksellian Point-Input, Point-Output Model of Capital Accumulation: A Modern View (or Neoclassicism Slightly Vindicated)," *J. Polit. Econ.,* Jan. 1973, *81,* 71–97.

R. W. Clower, "Is There an Optimal Money Supply?" *J. Finance,* May 1970, *25,* 425–33.

———, "Theoretical Foundations of Monetary Policy," in G. Clayton et al., eds., *Monetary Theory and Monetary Policy in the 1970's,* London 1971.

——— **and P. W. Howitt,** "Money, Credit, and the Timing of Transactions," disc. pap. no. 72, dept. econ., U.C.L.A. 1976.

M. R. Darby, "The Allocation of Transitory Income Among Consumers' Assets," *Amer. Econ. Rev.,* Dec. 1972, *62,* 928–41.

A. Dixit, J. Mirrlees, and N. Stern, "Optimum Saving with Economies of Scale," *Rev. Econ. Stud.,* July 1975, *42,* 303–25.

E. L. Feige and M. Parkin, "The Optimal Quantity of Money, Bonds, Commodity Inventories, and Capital," *Amer. Econ. Rev.,* June 1971, *61,* 335–49.

J. S. Flemming, "The Utility of Wealth and the Utility of Windfalls," *Rev. Econ. Stud.,* Jan. 1969, *36,* 55–66.

Milton Friedman, *A Theory of the Consumption Function,* Princeton 1957.

George Hadley and Murray C. Kemp, *Variational Methods in Economics,* Amsterdam 1971.

——— **and T. M. Whitin,** *Analysis of Inventory Systems,* Englewood Cliffs 1963.

C. Lluch, "Money, Commodity and Input Demand Functions," work. pap. 7101, Institut des Sciences Economiques, Université Catholique de Louvain, July 1971.

H. Lorie, "An Inventory Theoretic Approach to the Transaction Demand for Money, and the Intertemporal Consumption Allocation," paper presented at meetings of the Econometric Society, Toronto, Dec. 1972.

J. M. Ostroy, "The Informational Efficiency of Monetary Exchange," *Amer. Econ. Rev.,* Sept. 1973, *63,* 597–610.

D. D. Purvis, "Short-Run Dynamics in Models of Money and Growth," *Amer. Econ. Rev.,* Mar. 1973, *63,* 12–23.

M. Sidrauski, "Rational Choice and Patterns of Growth in a Monetary Economy," *Amer. Econ. Rev. Proc.,* May 1967, *57,* 534–44.

J. Tobin, "The Interest Elasticity of the Transactions Demand for Cash," *Rev. Econ. Statist.,* Aug. 1956, *38,* 241–47.

H. Uzawa, "Time Preference, the Consumption Function and Optimum Asset Holdings," in J. N. Wolfe, ed., *Value, Capital, and Growth,* Chicago 1968.

Did the 1968 Surcharge Really Work?: Comment

By ARTHUR M. OKUN*

In a recent issue of this *Review*, William Springer arrives, in his words, "at exactly the opposite conclusions" (p. 644) to those of an article I wrote concerning the effects on aggregate consumer expenditures of the 1968–70 temporary income tax surcharge.

As Springer correctly states at the outset of his article, I performed my experiment "using the consumption sector equations from four prominent macroeconomic models (Data Resources, Inc. (*DRI*), Wharton, Office of Business Economics, and Michigan)" (p. 644). In view of the occasional subsequent references to "Okun's equations" (pp. 645, 656), let me make clear that I did not fit or refit any equations in performing my experiment. In order to avoid any contamination of the result by decisions on specification that I might make, I relied entirely on equations developed by experienced model builders who had formulated their consumption sectors with no test of the tax surcharge in mind. Thus, when Springer complains that "Okun's approach suffers from its lack of grounding in a well-developed theoretical framework . . ." (p. 645) or when he chides me for not using "properly specified equations" (p. 657), the real targets of his strictures are Otto Eckstein, Saul Hymans, Lawrence Klein, Maurice Liebenberg, and their colleagues. I am delighted to be placed in such outstanding company.

The consumption of nondurables and services occupies the center ring in my postmortem of the tax surcharge and in Springer's critique of it. That is where the analytical case of permanent-income theorists argues against the effectiveness of the tax surcharge. According to their hypothesis, a change in measured disposable income produced by a *temporary* tax surcharge would be far less effective in restraining consumption of nondurables and services than would a permanent or standard drop in income of equal size. The four models I used in my experiment deliver a clear-cut verdict on that issue. Based on their prediction errors in a dynamic simulation for 1968:3 to 1970:3, three of them yielded estimates that the surcharge reduction in disposable income was at least as effective per dollar as a standard cut in income in curbing demand for nondurables and services. The fourth (*DRI*), implied that the tax surcharge was 69 percent as effective per dollar as a standard change in disposable income. As I concluded, " . . . the data provide no reason for questioning the effectiveness of the surcharge on those components of consumption where the basic challenge of the permanent income hypothesis was focused" (p. 192).

In his article, Springer develops his preferred equation to explain the consumption of nondurables and services (including the current services of durables). That equation does indeed deliver a sharply contrasting verdict that the tax surcharge was (at most) zero percent effective in curbing consumer demand. Springer implies that the evidence from his equation simply swamps the conflicting evidence from the consumption sectors of the four econometric models. In a concluding discussion of the effectiveness of flexible changes in income taxes, he even states that "This paper has attempted to lay this myth to rest once and for all . . ." (p. 658).

To make such a claim, one must have great conviction that he has built a better mousetrap. Yet Springer offers no evidence whatsoever of how well his mousetrap performs in comparison with its rivals. The reader is given no informa-

*Senior Fellow, Brookings Institution. The views expressed are my own and are not necessarily those of the officers, trustees, or other staff members of the Brookings Institution.

tion on whether Springer's equation fits the sample period better than the consumption sectors of the models, whether it extrapolates better to the surcharge period, or whether it is more consistent with observations from other sources such as cross sections of households or time-series of other countries. To the best of my knowledge, Springer's article is unique in economic literature in purporting to resolve an empirical dispute among alternative equations without recourse to empirical evidence on their comparative performance.

The missing empirical link is regrettable for reasons that transcend the issue of the tax surcharge. It is unfortunate, in general, that theory and econometric practice in the field of aggregate consumer demand have drifted apart. Most theorists support the permanent income (or long-horizon or life cycle) hypothesis—here, Springer is in good company. Yet, most econometricians continue to rely on formulations with fairly short lags because they find them to work better empirically, particularly in capturing the cyclical swings in consumption. Reconciling that difference would be a genuine achievement. But any permanent income econometrician who seeks to do that must take on the task that Springer eschewed—proving that the long-horizon formulation fits the facts better than its rivals.

Because of minor differences among the equations (in the measurement of consumption and in the sample periods), I cannot offer a definitive comparison to fill in the missing empirical link. But a quick look at the statistics Springer presents on his own equation does not lead me to embrace it in preference to those of the models. His equation (17) has a standard error of $1.4 billion (1958 prices). The consumption sectors of the four models I used had standard errors ranging from $1.2 to $1.5 billion. In a dynamic simulation of the surcharge period of 1968:3 to 1970:2, his preferred zero-effect view has a root-mean-square error of $2.1 billion, higher than the *DRI* 1.0, Michigan's 1.6, and a shade above Wharton's 2.0, although lower than the Office of Business Economics figure of 3.3 (all based on my preferred full-effect view).[1] So far as I can tell the empirical performance of Springer's equation is not bad; but it seems a little below the average of the four models, and it is certainly not a super equation.

The strength of Springer's confidence in his equation reflects the strength of his priors. As he sees it, the equation is ". . . derived precisely from standard economic theory . . ." (p. 652). That consumption equation has three independent variables: 1) disposable income of the current quarter; 2) expected disposable income constructed with an adaptive expectations mechanism applied to past actual disposable income; and 3) a term which multiplies expected disposable income by the real interest rate (constructed with an adaptive expectations mechanism applied to past inflation rates). Its theoretical properties raise some questions in my mind. Does "standard" economic theory exclude wealth as a variable in the consumption equation? Does it call for an interest rate to be interacted with expected income? Does it leave one

[1]In his reply (following), Springer takes issue with this comparison, arguing that there are two equally valid ways of comparing the performance during the surcharge period, using either full-effect or zero-effect simulations. But, I submit, my simulation comparison is the only way to throw light on the single key empirical question: what set of consistent beliefs would have given the best predictions for the surcharge period? For that purpose, the simulations must consistently use the full-effect view in the fitting and simulation of the equations that support the full-effect view and must consistently use the zero-effect view in the fitting and simulation of Springer's equation, which supports zero effect. Once the models tell me to believe full effect, I do not care how poorly they perform on the zero-effect view. Nor would I fault Springer's equation, which supports zero effect, for failing to track consumption on the full-effect view. As reported above, with that consistent test, Springer's equation finishes fourth in a field of five—a result that, while not decisive, is appropriate evidence and not "misleading."

On the other hand, Springer's demonstration that a bad equation can "duplicate Okun's finding" and make the surcharge look good is not relevant adverse evidence on the equations from the econometric models. Although Springer's "how-not-to-do-it" equation has a few of the characteristics of the consumption equations of the models, it has enormous differences, as evidenced by a standard error twice as large as those of the models! The consumption equations from the four models, which also make the surcharge look good, are good equations; and their verdict cannot be dismissed.

entirely comfortable with Springer's empirical result that a rise in real interest rates actually increases current consumption? Does standard permanent income theory identify temporary tax changes as the sole source of transitory income, leaving receipts from working overtime, year-end extra dividends, and the like in the permanent component? Does standard long-horizon theory suggest that the impact of the current quarter's income on consumption will be as large as the 0.26 coefficient estimated by Springer? Does it imply—as Springer's equation does—that even a permanent tax change like that of 1964 would be plugged into the expected income of the consumer at the rate of only 10 percent per quarter?

In fact, there is no such animal as "standard economic theory," and neither Springer's equation nor any other one is derived precisely from it. Any adequate economic theory must solve the puzzle posed by a pattern of behavior discerned in a number of cross-section studies (noted in my article, pp. 176–78) that focused particularly on windfall income. They find that small windfalls, positive or negative, are not treated significantly differently by the consumer than is a standard change in income. On the other hand, large windfalls have substantially smaller effects on consumption than do standard changes in income, much as the permanent-income hypothesis would imply.

Those empirical findings suggest to me that people really do care about equalizing their living standards over the long run (in line with the permanent-income or life cycle view), but also that they recognize that the estimation of income over a long-term horizon is bound to be highly uncertain and very costly (a matter ignored by that view). It seems reasonable that, because the required forecast of "permanent income" is so difficult and so unrewarding, people tend to gear consumption to it mainly when their take-home pay ("measured income") is clearly and substantially abnormal. In the event of a large positive windfall, they will not raise their basic living standard commensurately; in response to a substantial negative windfall, they will not tighten their belts very far. But when they get a little overtime pay, an extra dividend check, or a little dent in their income through a temporary tax surcharge, they may well not find it worthwhile to segregate that part of their receipts. Nor can they really estimate its permanence. Most Americans in 1968 knew that some of the Korean War "temporary" taxes of the early 1950's had lasted for more than a decade. They had no way to predict the duration of the Vietnam War tax surcharge, and I suspect that most had the good sense not to try. To develop point-estimate expectations of long-run disposable income, the household would have to apply varying coefficients of expectation to all sorts of actual changes in income and to weigh all sorts of future contingencies. That complicated game is not likely to be worth the candle. Archie and Edith Bunker may have better uses for their time, and so they can live rationally with the view that they can afford to consume out of current income, in the absence of clearly contradictory information.

There are analytical as well as empirical pitfalls for any theory that would make a crisp distinction between permanent and temporary tax changes. Let me illustrate with the following example. Suppose that, late in 1975, the Congress had passed Act *A*—a large permanent tax cut to take effect at the start of 1977. That act is identical to a two-part Act *B* consisting of 1) a permanent reduction starting in 1976, and 2) an equal temporary increase for 1976 alone. If the temporary increase in *B2* exerts *no* restraint on consumer demand, then *A* must have exactly the same effect on consumption in 1976 as would *B1* taken alone. The permanent tax cut starting in 1977 must be just as stimulative in 1976 as the one starting in 1976. Does anyone believe that proposition? Anybody who does not is admitting the effect of temporary income tax changes.

Two separate points raised by Springer's article—concerning autos and the saving rate—also require brief comment. Let me reiterate my puzzlement about the behavior of automobile demand in the period from mid-1968 to the end of 1969. It was simply far stronger than would

have been predicted by any of the automobile equations of the models, even assuming that the surcharge had no restraining effect on that sector. Logically, one may believe any of the following propositions: 1) the surcharge actually stimulated the demand for automobiles; 2) the surge in automobile demand was made possible by a belt-tightening in nondurables and services, which, in turn, gave a misleadingly favorable picture of the effectiveness of the surcharge on those components in the equations I used; 3) the surge in the demand for automobiles (whatever its motive) was financed by a cut in the saving rate and would have taken place regardless of the surcharge. In my own judgment, I find the first proposition entirely implausible; while I cannot rule out the second, I know that particular strength or weakness in automobile demand is usually associated with lower or higher saving rather than with lower or higher levels of demand for other consumer goods and services (pp. 197–98). And so I rest (uneasily) on the third proposition. Given that alternative, the range of estimates of total effectiveness of the surcharge (compared to a standard change in income) for the four econometric models is, as I reported, 63 to 88 percent.

The low level of the saving rate observed during the surcharge period that Springer stresses reflects in part the surge in automobile demand. But more generally and more importantly, the saving rate is a misleading indicator of abnormally strong or weak consumer demand whenever the rate of growth of income fluctuates. Neither Springer's equation nor anyone else's generates a stable saving rate; with any lagged response in consumption, the predicted saving rate rises temporarily when income accelerates and falls temporarily when income decelerates —whatever the source of the fluctuation in income. When the growth of real income slowed markedly after mid-1968, a drop in the saving rate was to be expected.

In summary, Springer has developed an interesting new consumption equation, but he offers no empirical case for it relative to its competitors, and his analytical case is not persuasive. Viewing Springer's results charitably, I would conclude that the score in favor of the efficacy of the 1968–70 temporary tax surcharge in curbing nondurables and services now stands at four to one instead of four to zero.

REFERENCES

W. L. Springer, "Did the 1968 Surcharge Really Work?" *Amer. Econ. Rev.*, Sept. 1975, *65*, 644–58.

A. M. Okun, "The Personal Tax Surcharge and Consumer Demand, 1968–70," *Brookings Papers*, Washington 1971, *1*, 167–212.

Did the 1968 Surcharge Really Work?: Reply

By WILLIAM L. SPRINGER*

The evidence put forward by Arthur Okun still does not bear the weight of his conclusion that the 1968 surcharge effectively curbed expenditures on nondurables and services. In my original article I was able to show how his results depended on certain defects in formulation and econometric technique. Okun's reply does not address these issues and consequently my conclusion that the surcharge did not work remains intact.

I

Okun chides me for not resting my argument that the surcharge did not work on a comparison of the statistical properties of my equations with those that he used. But he then proceeds to admit that there is very little to choose among the equations since they all fit the data extremely well (largely because of the strong time trend in both the dependent and the independent variables). Moreover, he makes a misleading comparison between the root-mean squared errors (*RMSE*) of my zero-effect equation and his full-effect equation. If the object of the exercise is to "see which equation is best," then the estimation and simulation procedures should be made as comparable as possible so that any differences in error behavior will be the result only of alternative specifications. The proper comparison is between the *RMSE* of the zero-effect view in dynamic simulations with my equation, estimated using the *full*-effect definition of income, and Okun's zero-effect equations, all of which were also (incorrectly) estimated using the full-effect definition of income. The results are displayed in Table 1, and my equation clearly fits the data better than any of the forecasting equations used in Okun's analysis. On the other hand, one could also compare full-effect equations estimated with full-effect income, in which case Okun's equations come out on top (see Table 2).

TABLE 1—COMPARISON OF ZERO-EFFECT VIEW USING EQUATIONS ESTIMATED WITH FULL-EFFECT INCOME

Equation	Dynamic *RMSE* 68–III to 70–III
Springer	1.4[a]
Data Resources, Inc.	1.8
Michigan	3.7
Wharton	4.1
Office of Business Economics	5.7

[a] 1968–III to 1970–II

TABLE 2—COMPARISON OF FULL-EFFECT VIEW USING EQUATIONS ESTIMATED WITH FULL-EFFECT INCOME

Equation	Dynamic *RMSE* 68–III to 70–III
Springer	3.9[a]
Data Resources, Inc	1.0
Michigan	1.9
Wharton	2.2
Office of Business Economics	3.6

[a] 1968–III to 1970–II

The purpose of the foregoing is merely to illustrate that it is not possible to arrive at a definitive answer as to which is the "best" equation, by looking at the *RMSE*, or even at $\bar{R}^2$, *D.W.*, etc. as well. The statistical differences among the equations are unlikely to be great enough or consistent enough to permit such a determination.

But more importantly, the issue between Okun and myself is only partly one of equation specification. The key question is why did Okun and I get opposite results performing the same experiments with slightly different equations. It is therefore necessary to compare the procedures used in making these tests. If the discrepancy cannot be explained except as a consequence of using different equation specifications

*Senior staff economist, Council of Economic Advisers.
I have benefitted from correspondence with Arthur Okun

(as Okun seems to think), and if all the equations perform about as well statistically, then indeed it is my priors against his and the objective verdict stands at four-to-one. On the other hand, if it can be shown that Okun's findings stem as well from errors in econometric technique, then the conclusion that his equations showed that the surcharge was effective in curbing expenditures on nondurables and services does not stand up.

This is the issue I discussed at some length in Section IVc of my paper, and which Okun has not addressed in his reply. There I was able to show that my results did not depend on the particular specification of the consumption function which I used. By not correcting for autocorrelation, by using a separately estimated equation to test the zero-effect view, and by dropping the real interest rate term I was able to duplicate Okun's finding that the full-effect view was superior. I concluded that "we may speculate that had Okun used properly specified equations which included a real interest rate variable, had he made the appropriate correction for autocorrelated errors, and had he simulated the two hypotheses with equations estimated using the corresponding definition of income, he could not have escaped the conclusion that the 1968 surcharge did not effectively reduce consumer expenditures" (p. 657). Until Okun can show that he still gets the same results when he uses the correct econometric procedures with his forecasting equations, his claim that the 1968 surcharge was a success remains unproven.

II

On a more general level, I noted in my original article that Okun's paper suffered from its lack of grounding in a well-developed theoretical framework. This was meant as a reminder that it is difficult to know what has been proven or disproven by a particular bit of evidence without reference to theory. As any good applied econometrician knows, empirical analysis can be used to "prove" just about anything. The role of theory is to guide the analyst in designing the correct experiments and to indicate how the results should be interpreted. A good researcher never accepts his results at face value, but rather evaluates them in relation to a set of first principles, and asks whether or not they make sense.

The theoretical model relevant to this particular episode of fiscal policy is the permanent income/life cycle hypothesis of consumption. Okun's rhetorical questions notwithstanding, this *is* in fact the "standard" theory of aggregate consumer behavior and quite obviously lies behind the forecasting equations used by Okun. The fact that their specifications differ from mine has more to do with the exigencies of putting together operational forecasting models than with any obvious disagreement on theory between myself and the four eminent economists invoked by Okun. Of course, the theoretical model from which one departs depends on the strength of one's priors. In this case, Okun expresses reservations about the permanent income hypothesis but provides no alternative model to replace it. In particular, he does not give a convincing rationale as to why consumers *should* have been affected by the surcharge, other than suggesting that consumers may not be up to the job of correctly optimizing over their utility functions.[1] His conclusion thus stands in a vacuum with no explanation as to why it is plausible.

By not basing his analysis firmly within the structure of the permanent income hypothesis, Okun runs into two problems. In the first place, as I noted in my original paper (fn. 7), Okun's experiment is not a proper test of the validity of the permanent income hypothesis in this particular instance. The experiment is designed to show whether or not the surcharge "worked," and the results can only be consistent or incon-

[1] By relaxing one of the assumptions of the permanent income hypothesis—viz., that capital markets are perfect—it is possible to show that a temporary tax may have a powerful first-round impact on spending and that the marginal propensity to consume *(MPC)* in this case will vary with the size of the tax. However, the conditions under which such a phenomenon might occur are not likely to exist in the aggregate (see my dissertation for an elaboration of this point). The notion of a "size effect"—whereby the permanent income hypothesis is somehow temporarily suspended and consumers treat large and small windfalls differently—is not very plausible.

sistent with the theory. They will not in any sense prove or disprove it. Secondly, Okun should have been more skeptical of his results because they were clearly at odds with the theory. This in turn might have led him to find the errors that I did.

III

One final note on using equations from large macroeconometric models to test hypotheses in single equation experiments. As I said earlier, such equations are designed to be well behaved in full-dynamic simulations within the context of the complete model. This means that these equations will often be specified differently than if they had been derived exactly from first principles to test a particular hypothesis. And as any model builder knows, the behavior of individual equations is often quite different in single equation and complete model simulations. Consequently, care should be exercised in concluding that a model says this or that based on an experiment which extracts a single equation and simulates it alone. Moreover, the consumption equations in these models are undoubtedly specified differently today than at the time Okun wrote his paper, and could very well give different answers. This is yet another reason to start from a firm theoretical base when examining issues such as the surcharge.

REFERENCES

A. M. Okun, "The Personal Tax Surcharge and Consumer Demand, 1968–70," *Brookings Papers*, Washington 1971, *1*, 167–212.

———, "Did the Surcharge Really Work?: Comment," *Amer. Econ. Rev.*, Mar. 1977, *67*, 166–69.

W. L. Springer, "Did the 1968 Surcharge Really Work?," *Amer. Econ. Rev.*, Sept. 1975, *65*, 644–59.

———, "Windfalls, Temporary Income Taxes, and Consumption Behavior," unpublished doctoral dissertation, Princeton Univ. 1974.

Local vs. National Pollution Control: Note

By FREDRIC C. MENZ AND JON R. MILLER*

In recent issues of this *Review*, several authors have debated the merits of various pricing strategies for the improvement of environmental quality.[1] While much of the exchange was merely the result of confusion surrounding the notation in Jerome Stein's comment on the 1971 Report of the President's Council of Economic Advisors, a fundamental issue still remains. Stein argued that in order to achieve economic efficiency in pollution abatement, the charge (or price) per unit of pollution damage must be the same in all localities and, to assure this result, pollution control pricing decisions should not be the responsibility of local government.[2] The question to be discussed here is not which level of government has the better technical expertise or information to accurately assess these pollution prices, but whether uniform nationwide prices for pollution damage are necessary for economic efficiency.[3]

In reply to criticisms of his argument (particularly those of Lerner), Stein suggested that his results follow from his assumption that residents of one area are concerned with the level of environmental quality in all other areas, whereas Lerner's result (different prices for pollution damage among localities) would follow if the level of environmental quality in a local area is of concern only to the residents of that particular area.

The purpose of our note is to show that Stein's result is based on a rather restrictive specification of the social welfare function that does not allow for differences in regional preferences for environmental quality. If such differences exist, nationwide *uniformity* of prices for pollution damage is unlikely to be efficient, even when environmental quality is a "national public good."

Assume there are two regions where Y_i represents output, X_i are the pollution residuals, and x_i is the pollution damage ($x_i = H(X_i)$) in region i.[4] Output is related to pollution damage through a general transformation

$$(1) \qquad Y_i = f_i(x_i)$$

where $f'(x_i) > 0$ and $f''(x_i) < 0$.

If the effects of pollution in a region are confined within its own borders, then the local utility function for the representative individual in each region is

$$(2) \qquad V^i = V^i(Y_i, x_i) = V^i[f_i(x_i), x_i]$$

If each of the two regions maximizes the utility only of its own inhabitants (i. e., there are no utility interdependencies between the two regions), the social welfare function may be formulated

$$(3) \qquad W^L = nV^1[f_1(x_1), x_1] + (1-n)V^2[f_2(x_2), x_2]$$

with the weights n and $(1-n)$ assigned according to population. Stein considers this function to be implicit in Lerner's analysis (the superscript L refers to Lerner). Welfare maximization under these conditions, as noted by Stein, requires that the marginal rate of substitution between output and pollution equal the marginal rate of transformation in each region *separately*. Consequently, there is no need for the price of pollution damage to be equal in the two areas.

Now suppose that environmental quality is a

*Assistant professors of economics, Clarkson College of Technology. Comments from Sam Peltzman and Robert J. Latham are gratefully acknowledged.

[1]See Allen Kneese, Abba Lerner, Sam Peltzman and Nicolaus Tideman, and Jerome Stein.

[2]See Stein (1971, pp. 533 and 535).

[3]In commenting on Stein's argument, Kneese, Lerner, and Peltzman and Tideman note that charges which reflect local pollution conditions could, in principle, be set either nationally or locally. The existence of a national price setter does not imply that prices will necessarily be uniform across the nation.

[4]The basic model is that of Stein (1971)

public good that transcends regional boundaries. The welfare of individuals in one region is affected by the level of environmental quality in other regions. Stein posits that the utility function for a representative individual in region i is now

$$V^i = V^i(Y_i, x_i, x_j) \tag{4}$$

To deal with this case, Stein formulates a social welfare function with the arguments composed of two weighted sums

$$W^S = W[nY_1 + (1-n)Y_2, nx_1 + (1-n)x_2] \tag{5}$$

or upon substitution,

$$W^S = W[nf_1(x_1) + (1-n)f_2(x_2), nx_1 + (1-n)x_2] \tag{6}$$

The prime minister of a hypothetical country that is comprised of the two regions maximizes aggregate social welfare by setting prices for pollution damage such that

$$f_1'(x_1) = f_2'(x_2) = -\frac{W_x}{W_y} \tag{7}$$

which results in identical prices for pollution damage in each region. Stein notes that this conclusion is a direct result of environmental quality being a national public good.[5]

The curious aspect of Stein's analysis is that in responding to Lerner's contention that extra-regional environmental quality is not an argument in local utility functions, he ignores the local utility functions altogether. Stein presents the utility function of an area's residents as equation (4), but this utility function is not incorporated into the social welfare function, equation (6). In Stein's formulation of Lerner's case, equation (3), the utility functions themselves are arguments of the prime minister's welfare function.

An alternative formulation, and one that seems more consistent with Stein's assumptions about the nature of environmental spillovers, would introduce interdependencies but allow individuals in the two regions to evaluate the spillovers differently. Viewed in this way, nationwide price uniformity, on the one hand, and local price differentials, on the other, become polar cases on a continuum based on the extent of environmental interdependence between the regions.[6]

Let the social welfare function be that which Stein used for the Lerner case, but now with spillovers between the two regions. The social welfare function becomes

$$W = nV^1[f_1(x_1), x_1, x_2] + (1-n)V^2[f_2(x_2), x_2, x_1] \tag{8}$$

To maximize aggregate welfare the prime minister would choose prices for pollution damage such that

$$f_1'(x_1) = -\left[\frac{\frac{\partial V^1}{\partial x_1}}{\frac{\partial V_1}{\partial Y_1}} + \frac{1-n}{n}\,\frac{\frac{\partial V^2}{\partial x_1}}{\frac{\partial V_1}{\partial Y_1}}\right]$$

$$f_2'(x_2) = -\left[\frac{\frac{\partial V^2}{\partial x_2}}{\frac{\partial V^2}{\partial Y_2}} + \frac{(n)}{1-n}\,\frac{\frac{\partial V^1}{\partial x_2}}{\frac{\partial V^2}{\partial Y_2}}\right]$$

With this formulation, there is no apparent reason why $f_1'(x_1)$ would equal $f_2'(x_2)$, and hence, why the price of pollution damage in the two regions should be the same. The only difference between these optimality conditions and those of Lerner's case is the second term in each of the brackets, which results from including the externality argument in each region's utility function.[7] This additional term repre-

[5]Note, however, that the charges for X—the actual effluent discharges—are likely to differ between the regions due to different damage functions.

[6]The analogy of this analysis with the public and private goods continuum (with pure public and pure private goods as polar cases) should be apparent. For a discussion of the public goods case, see E. J. Mishan, pp. 9–14, and James Buchanan, pp. 49–76

[7]Note that the size of this difference depends on the magnitude of the marginal interregional externality. As the $\partial V^i/\partial x_j$ approaches zero, the optimal prices for pollution damage approach those of Lerner's example. However, convergence of these two cases does not require a complete absence of interregional spillovers For example, the effect of region i's pollution on the residents of region j might be represented by a constant function, or one that is relatively flat over the relevant range. These conditions could exist if distance from the polluting region lessens the impact of marginal changes in pollution damage, i.e., residents of region j suffer losses in utility from the presence of pollution in region i, but these utility losses are constant over a broad range of pollution damage.

sents the ratio of the populations (or their weights) in the two regions multiplied by the marginal rate of substitution of pollution in one region for output in the other. So long as the interregional spillovers are taken into account, pollution damage would be less (as is the usual case when negative externalities are internalized), but there is no reason for the marginal rate of transformation between environmental damage and output to be identical in the two regions. Whether the pollution damage prices emanating from the maximization of this social welfare function would result in substantial relocation of population and economic activity is another question. It would depend on the relationship of the costs of industry relocation, population migration, and the magnitude of the differentials in pollution prices. This is a question of dynamics and the stability of the equilibrium, and is beyond the scope of this analysis. Note, however, that nationwide uniformity of prices for pollution damage, and the associated differentials in effluent or residuals charges, could still call forth movement of firms from areas where effluent charges are high to areas where they are low. No a priori judgment can be made regarding the ability to sustain the different equilibria. Resolution of these dynamic questions will require the formulation of models more elaborate than those employed by the contributors to this debate. All that has been shown here is that at any point in time the existence of environmental spillovers from one locality to another is *not* a sufficient condition for nationwide uniformity of pollution damage prices.

REFERENCES

James M. Buchanan, *The Demand and Supply of Public Goods*, Chicago 1968.

A. V. Kneese, "Pollution and Pricing," *Amer. Econ. Rev.*, Dec. 1972, *62*, 958.

A. P. Lerner, "Priorities and Pollution: Comment," *Amer. Econ. Rev.*, Sept. 1974, *64*, 715–17.

E. J. Mishan, "The Postwar Literature on Externalities: An Interpretative Essay," *J. Econ. Lit.*, Mar. 1971, *9*, 1–28.

S. Peltzman and T. N. Tideman, "Local versus National Pollution Control: Note," *Amer. Econ. Rev.*, Dec. 1972, *62*, 959–63.

J. L. Stein, "The 1971 Report of the President's Council of Economic Advisers: Microeconomic Aspects of Public Policy," *Amer. Econ. Rev.*, Sept. 1971, *61*, 531–37.

———, "Priorities and Pollution: Reply," *Amer. Econ. Rev.*, Sept 1974, *64*, 718–23.

Environment—Externalizing the Internalities?

By ABBA P. LERNER*

The debate on pollution charges between Jerome Stein and myself (1971, 1974) involved a minor and major issue. The minor issue was concerned with the distinction between pollution and pollution damage. The major issue was whether the charge per unit of pollution damage should be set separately for each locality where it is perpetrated (my position) or should be uniform throughout the economy (Stein's position).

In his last contribution (1974), Stein claims that "the main difference between us stems from [differences in] our implicit social welfare functions" (p. 718). He agrees that I would be right if "it is only the local population that feels the damage" (p. 720), but explains that his statement that "the environment belongs to the United States as a whole" was intended to mean that ". . residents of other areas and future generations have an interest in the environment of the *i*th [i.e., the local] region . . . This is the crux of the difference between us" (pp. 720–21). On the minor issue he repeats the agreed proposition that the charge should be based on "pollution damage" and not on "pollution or emission." He concludes with the "hope that the implicit assumption, ambiguities and differences between Lerner and myself have been clarified" (p. 723).

Fortunately this did not resolve our differences. I say fortunately, not unfortunately, because Stein's clarification has made me aware of a much more fundamental basis for our disagreement. What Stein calls "the main difference" turns out to be much less important than the minor issue. Both hold to the principle that the charge should take into account the damage to *everybody affected* by an externality, whether he lives in one region or another. If Stein's statement that "the environment belongs to the United States as a whole" means that *everybody* (at least in the United States) is in fact damaged by every externality, it is an *empirical* assertion. If this is the case the charge should indeed be equal to the *total* national (or universal) damage, i.e., including the damage *external* to the locality. But to sanctify such an empirical fact as implicit in the social welfare function turns it into a dangerous dogma.

Since there is agreement on the principle, the remaining difference would seem to be only about *administration*—about whether the highest authority should figure out the actual charge itself or whether it should allow the charge to be set locally, only making sure that damages external to the locality are also properly taken into account. This decision would have to be based on judgements of the relative administrative capabilities and difficulties and the relative efficiency and honesty of local and central governments, and on whether the different regions can charge each other for pollution spillovers.

There remains the puzzle that in different regions the same pollution damage, as defined by Stein, causes different degrees of loss of utility and should therefore be discouraged in different degrees by differing charges. Yet Stein proves that the overall efficiency principle of proportionality of price to marginal cost requires the charge for the same pollution damage to be *the same* throughout the economy.

The solution to this puzzle—the true crux of the difference between us—lies in our definitions of "pollution damage." I take pollution damage to mean utility losses for which those who suffer them are not paid or compensated (which is what makes them externalities). The appropriate charge is the amount of money that would just compensate the victims for the utility losses (disregarding any social damage or benefit from the redistribution of income or wealth). Stein tells us that his pollution damage "is not

*Queens College and the Graduate Center, City University of New York; and Florida State University. I am in debt for useful criticisms by my colleagues Michael Dohan and Donald Gordon.

measured in terms of utility but in the environmental argument of the utility function . . . The argument in the utility function is the purity of the water *(x)*, not the quantity of chemical dumped *(X)* . . . The 'loss of a unit of the environment' is precisely what I meant by a unit of *x*" (p. 719). Stein calls for a uniform national charge, not indeed per unit of pollution emission but per unit increase in the impurity of water. Stein's "pollution damage" is thus a *physical* change—a physical argument in the utility function.

This gives rise to many bothersome questions. What *is* the same increase in impurity at different levels of previous impurity? Does the volume of water affected come into the picture? What about different degrees of impurity at the beach, or at the pier end, or in midstream? On the surface? At various depths? etc., etc.

All these annoying questions can be dismissed if there is only one source of pollution and only one place where the increase in impurity of the water (or any other external effect of any perpetration of pollution) impinges on the public welfare. This happens to be the case in Stein's mathematical treatment of the problem in terms of a universe of two regions, both affected by one pollution with only one pollution damage in only one of the regions. But in this case there is no meaning to the proportionality of price to marginal cost. Only when we have more than one source producing "the same pollution damage" (the loss of an equal number of "units of environment") in more than one place, is there any meaning to equality of charge But then the nasty questions reappear and there is no reason for assuming that the different physical units of pollution damage are equally objectionable and should therefore be equally discouraged by equal charges.

The economic efficiency principle of proportionality of price to marginal cost (which inspires Stein's paper) derives from the more general principle of equality of marginal social benefit to social marginal cost (the alternative marginal social benefit foregone)—all in terms of utility. Equality of price or charge is appropriate for different units of the same input (or output).

This is also what would automatically come about in a perfectly competitive market if the units are perfect substitutes for each other and can be transported costlessly from place to place (or, alternatively, if all complimentary and substitute commodities, as well as the consumer beneficiaries or victims, can costlessly be moved).

Any inequality of charge would then indicate a failure to maximize the excess of output over input by failing to move outputs from where they are cheaper to where they are dearer until their prices and marginal utilities are equal everywhere. But if the different physical units of environment lost—the arguments in the utility function—are not equally obnoxious everywhere and so are not perfect substitutes, the argument does not hold. The different physical units of environment cannot be treated as units of the same commodity.[1] Each separate act of pollution must be judged and charged according to the damage to the *utilities* of the people affected. No physical measure can be a substitute for this. The shift from emission *(X)* to units of environment lost *(x)* still leaves us with a *physical* item and can therefore bring us no nearer to a solution.[2]

[1] I cannot resist quoting my 1934 article "The Concept of Monopoly and the Measurement of Monopoly Power" " objects having the same physical characteristics are not the same goods if they are at different places . *In calling the same thing at different places different commodities we have rejected the criterion of physical similarity as a basis for classification of commodities* . physical qualities, spacial and temporal position, are irrelevant now that we have the ultimate criterion of substitutability at the margin It would perhaps be best to . break with the traditional usage by speaking of "units of accommodation" instead of units of commodities "[p 217]

[2] The (average) increase in impurity (*x*) multiplied by the volume of water affected is equal to the volume of emission (*X*) This may help to explain my failure to recognize that by "pollution damage" or *x* Stein did not mean the utility loss He did not mean the unit of *pollution emission* measured as it leaves the polluter's premises, but it was still a *physical* item measured somewhere along the way from these premises to the disutilities that constitute the externality. My inability to envisage any such physical measure as economically relevant led me to assume, wrongly, that he was occasionally slipping from my pollution damage (utility loss) all the way back to pollution emission (his *X*). His slip was only as far as the *secondary* physical measure (his *x*).

The trouble is the same as that of the socialist planners who tried to outflank the market by providing physical definitions of products in their instructions to the plant managers. The plant managers, trying to reach and exceed their quotas in the plan, continually outwitted the socialist planners until these abandoned their "materialism" and required *sales* instead of *physical output* in quotas. This enabled the consumer to show whether the product was right or not by his decision to buy or not to buy. The shift was from *objectively* defined "arguments in the welfare function" to the *subjective* preferences of the consumer—his utility or disutility as expressed on the market.

This reform could not be effectively introduced as long as scarcity was so extreme that everything could be sold. The planners could then exercise complete sovereignty in deciding what should be produced and what the prices should be. A similar suppression of consumer sovereignty is latent in the mystical declaration that "the environment belongs to the United States as a whole." Stein's "prime minister" would have to define the objective, physical units of environment to be given uniform prices no matter how un-uniform their marginal utilities in different places. An activity of mine that affects my neighbor, while external to myself, may be internal to my street, but as an externality declared to affect the environment which belongs to the United States as a whole it would have to have its price fixed by the prime minister.

On this line of thought there is no reason why the environment as a whole (which belongs to the people as a whole) should be protected only in the case of public goods or what economists have called externalities. Even an activity that affects none but the person or persons undertaking it can be declared to affect the environment as a whole and taxed, or even prohibited. What we could get is not an *increase* in consumer sovereignty by internalization of externalities but the diminution of consumer sovereignty, of privacy, and of freedom by the *externalization of internalities*.

I do not believe for a moment that Stein has any such malicious intentions. I do, however, see here an example of the dangers of letting the mathematics run away with the economics. I conjecture that if he had had to keep on repeating the clumsy "loss of a unit of environment" instead of the elegant x, he would not have slipped. The danger is that instead of mathematical economics, which after all should still be a branch of economics, we can then get ecomathematics—a branch of pure mathematics that finds its axioms in some propositions in economics (or perhaps in ecology).

REFERENCES

A. P. Lerner, "The 1971 Report of the President's Council of Economic Advisers: Priority and Efficiency," *Amer. Econ. Rev.*, Sept. 1971, *61*, 527–30.

——— "Priorities and Pollution: Comment," *Amer. Econ. Rev.*, Sept. 1974, *64*, 715–17.

——— "The Concept of Monopoly and the Measurement of Monopoly Power," *Rev. Econ. Stud.*, June 1934, *1*, 157–75; reprinted in William Breit and Harold Hochman, *Readings in Macroeconomics*, New York 1971, 207–23.

J. L. Stein, "The 1971 Report of the President's Council of Economic Advisers: Micro-Economic Aspects of Public Policy," *Amer. Econ. Rev.*, Sept. 1971, *61*, 531–37.

——— "Priorities and Pollution: Reply," *Amer. Econ. Rev.*, Sept. 1974, *64*, 718–23.

Market Structure and Product Varieties

By LAWRENCE J. WHITE*

The problem of relating the amount of product variety[1] to market structure is an interesting one, on both theoretical and policy grounds. Under constant returns to scale, it is easy to demonstrate that a competitive industry will offer the optimal range of product varieties. Since there are no cost penalties for small scale production, competitive producers will satisfy every taste, as long as effective demands can cover unit costs.

An important question is whether a profit-maximizing monopolist, also operating under constant costs, will also offer this same optimal range of variety. In a recent article, Kelvin Lancaster has argued that a monopolist will offer this optimal range of varieties and, hence, that market structure does not matter in this respect Lancaster's argument reinforces a similar result demonstrated earlier by Peter Swan.

This paper will argue that the Lancaster-Swan result is generally incorrect and holds true only when a monopolist can practice price discrimination and effectively force the buyers of some varieties of goods to pay higher prices than other buyers of those same varieties. If the monopolist cannot so discriminate, he will not in general offer an optimal range of varieties but instead will offer nonoptimal varieties to some consumers or may even refuse to provide any satisfactory varieties to these consumers.[2]

*Associate professor, Graduate School of Business Administration, New York University. I would like to thank Michael Rothschild and Richard W. Parks for helpful comments on an earlier draft.

[1]There is a minor terminological problem that should be clarified. By a particular variety of a good, we mean what might also be called a quality level. The latter, though, sometimes conveys a sense of higher and lower quality, whereas variety can also encompass red shirts as against blue shirts.

[2]Similar results, to a greater or lesser extent, can be found in the models developed by Richard Parks, by William James Adams and Janet Yellen, and by Michael Mussa and Sherwin Rosen. The Parks result, though, is limited to a durability context, and the Adams-Yellin and Mussa-Rosen models are restricted to the cases in which individuals buy only single units of goods.

I. The Model

The Lancaster argument is deceptively simple. He argues (Theorem 4, p. 580) that the monopolist is trying to extract the most revenue that he can out of a consumer and hence will want to provide that consumer with the particular variety of product that provides the consumer with the most satisfaction. This provides the maximum potential consumer surplus for the monopolist to try to capture. It would be foolish for the monopolist not to provide a particular variety that a consumer wanted. This should be true in his dealings with each consumer, and each will get the particular variety that he desires. Hence, the optimal range of varieties will be offered.[3]

This argument ignores the possibility that the offering of a particular variety to a particular consumer at a price that maximizes profits from that consumer may allow a different consumer, who is consuming a different variety and is providing higher profits to the monopolist, to switch to the first variety with a lower overall profit result for the monopolist. If the monopolist cannot price discriminate so as to prevent this from happening, he will offer the first customer a non-optimal variety that is also less satisfactory to the second customer (and hence less likely to induce a consumption switch) or, at the limit, he may refuse to offer the first customer any satisfactory variety at all.

To demonstrate this result, we can use the same model developed by Lancaster. In Figure 1, the two axes represent the two quality attributes Z_1 and Z_2 that define a particular good. The ratio $r = Z_2/Z_1$ defines the particular varieties of the good. The good, then, can vary in variety from having all Z_2 and no Z_1 to having no Z_2 and

[3]Swan's result, expressed in terms of determining when a monopolist and a competitive industry will offer a new product, seems to rest heavily on the specific demand formulation that he used.

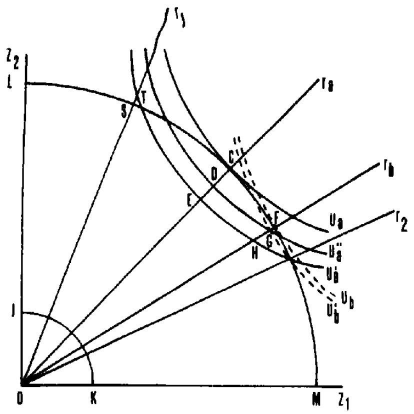


FIGURE 1

all Z_1, with all combinations in between. Movement outward from the origin along any ray, such as r_1 or r_2, implies increased amounts of the good of the same particular variety.

We can construct a product differentiation curve *(PDC)*, which shows the absolute amounts of the two attributes that can be produced from a given amount of resources devoted to the production of the good at the different quality levels. We have drawn one such curve *JK*, which can be defined as the amount of the two attributes that will be produced from a unit amount of resources. Since all points along *JK* represent the same amount of resources, they also represent equal costs and, under competition, all of the combinations of attributes along *JK* will sell for the same price. Larger amounts of inputs are represented by other curves, such as *LM*. By assumption, the *PDC* are homothetic expansions and exhibit constant returns to scale in the production of more goods from more resources. As drawn, the curves are concave to the origin, implying diminishing returns to the specialization of production on one attribute or the other. For geometric purposes, let us define the units of the attributes so that the product differentiation curves form perfect quarter circles.[4]

We can think of a consumer as having an indifference map defined over the two attributes of the good.[5] For the present purposes of geometry, we will assume that the consumer spends all of his income on the single good, and his problem is to choose the optimal quality level among the two attributes.[6] This is similar to the usual consumer choice problem, except attributes have replaced separate goods in the indifference function. Two different consumers' indifference maps, U_a and U_b, are represented in Figure 1.

Suppose that both consumers have the same income. Then, if a competitive industry is providing the goods to be purchased, this budget constraint can be represented by a *PDC* such as *LM*, since all points along *LM* represent equal cost combinations and hence equal revenue combinations. Consumer *A*, then, will choose the particular variety r_a and consume *OC* amount of goods and the appropriate amounts of the two attributes, while consumer *B* will choose the particular variety r_b. The competitive industry will, of course, provide both varieties in the market.

Now, suppose that a monopolist gains a franchise to be the exclusive producer of the good for all varieties between r_1 and r_2. Competitive firms can still produce all varieties to the left of r_1 and to the right of r_2. In considering the quality to offer to consumer *A* and the price to charge, the monopolist is constrained by the presence of the competitive firms offering potential substitutes beyond r_1 and r_2. Accordingly, the best that the monopolist can do is to try to extract the full income from consumer *A* but provide him with only *OE* amount of goods. This would reduce consumer *A*'s utility to U_a'; any further reduction of utility would cause consumer *A* to switch to purchasing a

[4]This allows us to make direct comparisons of distances along any ray from the origin, since, under competition, all the combinations at any distance will yield the same revenue. The general result, though, holds true for any linear or concave *PDC*.

[5]Like Lancaster, we assume that the good represents a noncombinable consumption technology, in which only one variety can be consumed at a time and attributes can be obtained only in those proportions represented by an available good; i.e., one cannot average over two varieties. See Lancaster, p. 572.

[6]One can easily construct mathematical examples that introduce a third dimension, all other goods and services, with the same results as those shown in the two-dimensional diagram.

particular variety just to the right of r_2 from a competitive firm. To achieve this, the monopolist charges a monopoly price that is OC/OE times the competitive price. His monopoly profits can be represented as distance CE. If consumer A's indifference map is homothetic, the same particular variety will be provided; if not, the particular variety will change but will still be optimal from consumer A's viewpoint, given the presence of the monopolist.

Similarly, the best that the monopolist can do with respect to consumer B is to charge a price that is OF/OG times the competitive price and provide him with OG amount of goods of variety r_b. Any higher price would cause consumer B to switch his purchases to the competitive firms beyond r_2. His monopoly profits are FG. Note that the monopoly price to B is lower than the monopoly price to A; $OC/OE > OF/OG$.[7] This is exactly what we would expect. Consumer B's tastes are closer to being optimally satisfied by the competitive firms on the fringe beyond r_2, and hence the monopolist is more constrained in what he can charge B.

But, if the monopolist actually does charge a price of OF/OG for quality r_b, then consumer A will find it worthwhile also to consume variety r_b, since he can then consume OG amount of goods which will yield him a higher level of satisfaction (U''_a) than the OE amount of variety r_a. The only way for the monopolist to prevent this is to charge A a price higher than OF/OH ($> OF/OG$) for variety r_b, which causes A to switch back to r_a. If the monopolist persists in charging OF/OG for r_b, then the best he can do is to drop the price of r_a to OC/OD ($< OC/OE$), which again causes A to switch back to r_a. Profits CD will always be greater than FG.

If the monopolist can practice price discrimination in selling r_b, charging a price of OF/OG to B and charging a price greater than OF/OH to A, then he will definitely offer both varieties in the market. But if he cannot discriminate, he will offer B an inferior variety—a variety that lies between r_b and r_2. At the limit, he may not offer B any satisfactory variety at all[8] (and B will simply purchase from the competitive fringe beyond r_2). As the monopolist offers a progressively inferior variety to B, his profits from B decline but his profits from A increase.[9] It is easy to show that the maximum profits for the monopolist must imply an inferior variety (or none at all) to B.[10]

The degree to which the profit-maximizing variety to B falls below the optimal level will depend solely on the shapes of the indifference curves for A and B and of the PDC. Or, if we go beyond the present geometry, it will depend on the technology of producing the varieties and on the numbers of the various consumers, their incomes, and their tastes relative to each others' tastes and relative to the particular varieties offered by the fringe of competitive firms. If most consumers are like consumer A and/or the consumer A individuals have higher incomes, the monopolist will find it increasingly worthwhile to provide B with an inferior variety or to ignore B entirely. A competitive industry would, of course, cater optimally to B's demands as well as to A's demands, regardless of numbers or incomes.

The only exception to this result occurs when B's tastes are so far away from A's that the provision of the optimal variety to B would not induce A to switch his purchases. This would occur, for example, if B's indifference curve

[7]Again, since the PDC curve is a quarter circle, these distances are comparable

[8]Or he can offer to sell r_b at a price above OF/OH, which is equivalent to not offering it, since neither A nor B will buy it at that price

[9]As the variety ray offered to B rotates from r_b toward r_2 (holding B's indifference map constant), distance FG (the profits from B) must diminish But the consumption of OG amount of goods of this inferior variety becomes less satisfactory to A (i e , U''_a decreases) and hence the monopolist can charge a higher price for r_a without inducing A to switch (i e , distance CD increases)

[10]This is proved as follows Let variety r_b^* be the ratio of Z_2/Z_1 actually offered to consumer B From fn 9 it is clear that the profits from B, Π_B, are a function of r_b^* and that the profits from A, Π_A, are also a function of r_b^* Overall profits, then are $\Pi(r_b^*) = \Pi_A(r_b^*) + \Pi_B(r_b^*)$ From fn 9, $d\Pi_A/dr_b^* < 0$ for all relevant ranges of r_b^* Also, $d\Pi_b/dr_b^* < 0$ for $r_b^* > r_b$, $d\Pi_b/dr_b^* = 0$ for $r_b^* = r_b$, and $d\Pi_b/dr_b^* > 0$ for $r_b^* < r_b$ Hence, at $r_b^* = r_b$, $d\Pi/dr_b^* < 0$, i e., at that point overall profits are not at a maximum, and the profit-maximization point must be at some point $r_b^* < r_b$ If that point is also $r_b^* < r_2$, then customer B will simply be ignored by the monopolist.

U_b were tangent to LM between S and T.

II. Conclusion

In general, then, market structure is likely to matter with respect to the range of product varieties. A monopolist may well find it in his interests to provide a nonoptimal product variety in response to minority tastes, or he may find it worthwhile to ignore these tastes entirely. A competitive industry, by contrast, would provide the optimal varieties. Only the ability to price discriminate would guarantee that the monopolist will provide the optimal range of varieties, and price discrimination may be difficult in practice.

One consequence of this result is that the welfare effects of monopoly may be underestimated by the usual method of measuring deadweight loss triangles for existing products. The presence of nonoptimal products or the absence of products from the marketplace, because of monopoly power exercised in the manner described above, may also yield a significant welfare loss that will simply not be measured in deadweight loss inferences from excess profit estimations.

REFERENCES

W. J. Adams and J. L. Yellen, "Commodity Bundling and the Burden of Monopoly," *Quart. J Econ.*, Aug. 1976, *90*, 475–98.

K. Lancaster, "Socially Optimal Product Differentiation," *Amer. Econ. Rev.*, Sept. 1975, *65*, 567–85.

M. Mussa and S. Rosen, "Monopoly and Product Quality," disc. pap. 75–12, dept. econ., Univ. Rochester, June 1975.

R. W. Parks, "The Demand and Supply of Durable Goods and Durability," *Amer. Econ. Rev.*, Mar. 1974, *64*, 37–55.

P. L. Swan, "Market Structure and Technological Progress: The Influence of Monopoly on Product Innovation," *Quart. J. Econ.*, Nov. 1970, *84*, 627–38.

Import Demand and Export Supply: An Aggregation Theorem

By RONALD W. JONES AND EITAN BERGLAS*

In basic trade theory the classical budget constraint links the value of a country's imports to the exports given up in exchange. If import demand is inelastic, a lower volume of exports is given up when import prices fall. Phrased differently, the lower volume of exports corresponds to a higher relative export price. It is as if the supply curve of exports is backward bending.

Whereas numerous examples have been given of commodities the import demand for which may be inelastic, it seems much harder to cite examples of commodities for which higher prices would fail to elicit greater exports. The standard textbook example involves the labor-leisure choice: A higher wage rate raises workers' real income and the demand for all (normal) commodities, such as leisure. If substitution effects are weak compared to the income effect of the wage rise, less labor would be supplied (exported) to the market. One reason it becomes harder to expect this result for produced commodities is that there is usually a substitution effect in production—a higher price encouraging greater output. A backward-bending offer or supply curve would entail that both this production effect and the substitution effect in consumption be slight compared with the effect of an export price increase in raising the real income of exporters and thus encouraging their own demand for the items they sell.

The purpose of this article is to resolve this difficulty by proving that even if the net supply curve for every commodity exported is positively sloped (in the sense that a higher price for any would elicit greater export of that item), a raise in all export prices could easily *reduce* the quantities exported. Furthermore, this systematic bias in aggregating exports does not apply to imports. As will be shown, this asymmetry in aggregation is attributable to the asymmetric impact of the income effect of price changes: net buyers are better off and sellers worse off when price falls.

*Professor, department of economics, University of Rochester, and professor, department of economics, Tel-Aviv University and University of Rochester, respectively

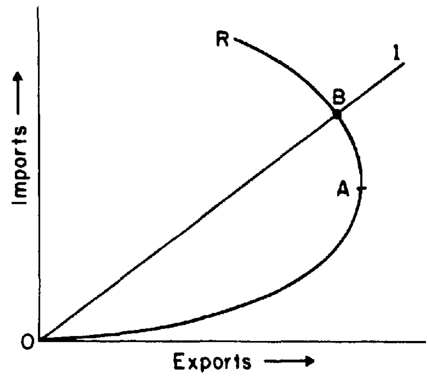


FIGURE 1

Figure 1 illustrates a country's offer curve as typically drawn. Import demand is elastic in range *OA* but, for lower import prices, becomes inelastic. For example, if line 1 shows initial terms of trade, a fall in relative import prices (or rise in the relative price of exports) would lower the quantity of exports since import demand is inelastic. Let ϵ denote the community's aggregate elasticity of demand for imports, defined as minus the relative change in import volume when the relative price of *all* commodities being imported rises by 1 percent, with *all* export prices constant.[1] Similarly, let γ denote the economy's aggregate elasticity of supply of exports, defined as the percentage increase in aggregate export volume when the price of each

[1]For the present ignore any nontraded commodities.

export commodity rises by 1 percent, with all import prices constant. The link between ϵ and γ must be given by (1) if imports are always paid for by exports, as assumed in Figure 1.

$$\epsilon - 1 = \gamma \tag{1}$$

Of course, in any relevant setting the import and export volumes shown by the offer curve are each composites of many traded items. Strictly speaking the rules of composites require all relative prices for commodities in a composite to remain constant. This we adhere to in Figure 1's offer curve. Equation (1) suggests that if imports are, on the aggregate, inelastic, the elasticity of export supply γ must be negative But what examples of negative export response to price can be provided? If it is difficult to cite examples, does (1) suggest as well that inelasticity of import demand would be rare? The aggregation theorem to be proved is that γ may be negative even though no individual export would fall in volume if that export price rises. That is, inelasticity in import demand and, by (1), a consequent fall in aggregate export volume as the terms of trade improve, is consistent with the absence of any commodity exhibiting a backward-bending supply response to a rise in its own price.

Aggregate import demand (ϵ) and export supply (γ) elasticities have already been defined. Now consider comparable expressions item by item. With a ˆ over a variable expressing relative changes ($\hat{x}$ is dx/x), let

$$\epsilon_i \equiv -\frac{\hat{M}_i}{\hat{p}_i} \tag{2}$$

where M_i is defined as the quantity of imports of commodity i and the price of i, p_i, is the *only* price assumed to change. Similarly, let X_j denote the export level of commodity j, and p_j its price. Then

$$\gamma_j \equiv \frac{\hat{X}_j}{\hat{p}_j} \tag{3}$$

again assuming that only p_j changes.

The obvious task is to relate aggregate elasticities (ϵ and γ) to elasticities for individual commodities (ϵ_i and γ_i). But note that the setting is different: ϵ expresses the aggregate response of import volume when *all* import prices rise, while ϵ_i captures the effect only of a rise in one import price. A similar remark can be made about γ. To bridge the gap we introduce expressions $\bar{\epsilon}_i$ and $\bar{\gamma}_i$ to designate the effect on any one item imported or exported when it is assumed that *all* import (or export) prices rise in the same proportion.

There is a simple relationship between aggregate ϵ and the $\bar{\epsilon}_i$. Suppose for convenience all prices are initially units so that $M = \Sigma M_i$. Since the assumption about prices is the same in the definition of ϵ and $\bar{\epsilon}_i$,

$$\epsilon = \sum_i \frac{M_i}{M} \bar{\epsilon}_i \tag{4}$$

That is, ϵ is a weighted average of all individual $\bar{\epsilon}_i$ In a similar fashion it is clear that.

$$\gamma = \sum_j \frac{X_j}{X} \bar{\gamma}_j \tag{5}$$

It is customary to note that each elasticity of import demand (or export supply) can be decomposed into the three reactions to a price change. a substitution effect in consumption; a substitution effect in production; and an income effect due to the price change.

Define the following substitution terms.

$$E_i^k \equiv \frac{\hat{D}_i}{\hat{p}_k} \tag{6}$$

where D_i is the total demand for commodity i, p_k is the only price to change, and real incomes are held constant. Each *own* elasticity, E_i^i, must be negative. Furthermore, homogeneity requires that $\Sigma_k E_i^k$ equal zero, letting k run over *all* commodities. If commodities are substitutes, E_i^k will be positive for $k \neq i$. In production let

$$e_j^k \equiv \frac{\hat{x}_j}{\hat{p}_k} \tag{7}$$

where the volume of production of commodity j is x_j and p_k is the only price to change. Own production terms, e_j^j, must all be positive while cross production terms can have either sign, but

will be negative in the case of substitutes. In any event, homogeneity requires $\Sigma_k e_j^k$ to be zero, summing over *all* commodities.

Now consider the decomposition of ϵ_i. The volume of imports M_i is the difference between demand D_i, and production x_i. Therefore

$$(8)\qquad \epsilon_i = -\frac{D_i}{M_i}\frac{\hat{D}_i}{\hat{p}_i} + \frac{x_i}{M_i}\frac{\hat{x}_i}{\hat{p}_i}$$

where p_i is the only price to change; D_i can be expressed as a function of all prices and real income. If only the ith price changes,

$$(9)\qquad \hat{D}_i = E_i^i \hat{p}_i + \frac{1}{D_i}\frac{\partial D_i}{\partial y} dy$$

where dy is the change in real income, measured in units, say, of some exportable (p_x). As p_i rises, real income is reduced by the terms of trade effect. That is,

$$(10)\qquad dy = -\frac{p_i M_i \hat{p}_i}{p_x}$$

The marginal propensity to import (or consume) commodity i, m_i, is defined as $(p_i/p_x)(\partial D_i/\partial y)$. Therefore

$$(11)\qquad \hat{D}_i = \left\{E_i^i - \frac{M_i}{D_i} m_i\right\} \hat{p}_i$$

The production of commodity i, x_i, depends on all prices, so that with only p_i changing,

$$(12)\qquad \hat{x}_i = e_i^i \hat{p}_i$$

Putting these together,

$$(13)\qquad \epsilon_i = \frac{D_i}{M_i}(-E_i^i) + \frac{x_i}{M_i} e_i^i + m_i$$

A similar procedure can be used to decompose the elasticity of export supply to yield

$$(14)\qquad \gamma_j = \frac{D_j}{X_j}(-E_j^j) + \frac{x_j}{X_j} e_j^j - m_j$$

The obvious difference, in comparing with ϵ_i, is that in export supply the income effect runs counter to the substitution effects. As export price rises, more is produced, less is demanded for any given level of real income, but real income rises, and part of this, m_j, spills over into increased home demand for the exported commodity. A strong enough income effect yields a backward-bending export response to a price rise.

Suppose we now consider how the response of imports or exports of a particular item varies when *all* prices in that category (i.e., all import prices or all export prices) change in the same proportion. On the import side $\bar{\epsilon}_i$ is given by expression (8), reinterpreted so that $\hat{p}_i$ has the same value for all import commodities i. The breakdown of $\hat{D}_i$ is now shown in (15):

$$(15)\qquad \hat{D}_i = E_i^i \hat{p}_i + \sum_{\substack{k\neq i\\ k\in M}} E_i^k \hat{p}_k - \frac{M}{D_i} m_i \hat{p}_i$$

where $\hat{p}_k = \hat{p}_i$ for all k in the class of imports. With all import prices rising in the same proportion ($\hat{p}_i$),

$$(16)\qquad dy = -\frac{p_i M}{p_x}\hat{p}_i$$

Note that the income effect is much more powerful than shown by (10) for the case in which only one import price rises The leverage on real income is given by aggregate imports M. The production response is shown by (17):

$$(17)\qquad \hat{x}_i = e_i^i \hat{p}_i + \sum_{\substack{k\neq i\\ k\in M}} e_i^k \hat{p}_k$$

Expression (18) for $\bar{\epsilon}_i$ puts these together:

$$(18)\qquad \bar{\epsilon}_i = \frac{D_i}{M_i}(-E_i^i) - \frac{D_i}{M_i}\sum_{\substack{k\neq i\\ k\in M}} E_i^k + \frac{x_i}{M_i} e_i^i + \frac{x_i}{M_i}\sum_{\substack{k\neq i\\ k\in M}} e_i^k + \frac{M}{M_i} m_i$$

The task of comparing $\bar{\epsilon}_i$ with ϵ_i is made more simple by rewriting (18) as:

$$(19)\qquad \bar{\epsilon}_i = \epsilon_i - \left\{\frac{D_i}{M_i}\sum_{\substack{k\neq i\\ k\in M}} E_i^k - \frac{x_i}{M_i}\sum_{\substack{k\neq i\\ k\in M}} e_i^k\right\} + \frac{(M - M_i)}{M_i} m_i$$

Two types of corrections must be made to ϵ_i to convert it into $\bar{\epsilon}_i$: (i) $\bar{\epsilon}_i$ will tend to be smaller than ϵ_i to the extent that rises in prices of all imports other than i will divert demand back to i

and production away from i. This is the expression in brackets, which we assume positive (commodity i is assumed to be a substitute for the group of all other importables). (ii) $\bar{\epsilon}_i$ will tend to be larger than ϵ_i to the extent that it incorporates a larger income effect. The $[(M - M_i)/M_i]m_i$ term reveals that with real income being reduced much more than if only one p_i rises, the effect on imports of good i (as captured by $\bar{\epsilon}_i$) is more drastic. These two types of corrections run counter to each other. With no further information as to how they can be compared, it is hard to tell if $\bar{\epsilon}_i$ is larger or smaller than ϵ_i. Therefore the aggregate elasticity of demand for imports, as shown in (4), may be larger or smaller than the average of all individual import demand elasticities, $\Sigma\,(M_i/M)\,\epsilon_i$.

A similar procedure holds, up to a point, on the export side. Suppose all export prices rise by the same percentage, with all import prices constant. The impact on the export supply of commodity j is, by definition, given by $\bar{\gamma}_j$. With the same decomposition as employed for imports, $\bar{\gamma}_j$ is related to γ_j by (20):

$$(20)\qquad \bar{\gamma}_j = \gamma_j - \left\{ \frac{D_j}{X_j} \sum_{\substack{k \neq j \\ k \in X}} E_j^k - \frac{x_j}{X_j} \sum_{\substack{k \neq j \\ k \in X}} e_j^k \right\} - \frac{(X - X_j)}{X_j} m_j$$

Each $\bar{\gamma}_j$ is related to γ_j, but again with two types of correction First, the fact that all export prices rise instead of just p_j serves in general to mitigate the substitution effects in consumption and production of the rise in the j price. That is, if all export prices rise, the cut in demand for j as of constant real income and the expansion in production of j will both be less pronounced than if j rose in price alone. But the income effect must be considered as well. This income effect on the export side serves to *raise* local demand for exportables and thus even further reduce export supply. That is, *both* types of correction serve to lower $\bar{\gamma}_j$ unambiguously below the γ_j value. Since the aggregate elasticity of export supply γ is a weighted average of the $\bar{\gamma}_j$, it systematically has a lower value than the typical export supply elasticity. Indeed, even if *all* γ_j are positive—not a single export good has a backward-bending supply curve—the aggregate γ could be negative. Inelasticity in import demand could be prevalent (so that ϵ could be less than unity), implying as well that γ must be negative. But this does not necessarily mean that any underlying γ_j export supply elasticity need be negative.

The asymmetry of the aggregation bias between imports and exports rests entirely upon the income effect. Compare (19) and (20). In both cases the adjustment in substitution effects to account for the rise in prices of other commodities (in the import or export group) serves to lower the value of $\bar{\epsilon}_i$ (or $\bar{\gamma}_i$) compared with the own effects, ϵ_i (or γ_i) But with all prices changing in a group instead of just one, the income effect is more powerful. On the import side the reduction in real income serves to raise the value of $\bar{\epsilon}_i$ compared with ϵ_i since import demand is reduced more by a larger fall in real income On the export side the supply to foreigners tends to be cut more the greater the rise in real incomes. On the import side substitution and income effect adjustments run counter to each other; on the export side they reenforce a value of overall γ that is lower than the average of the individual γ_j.

The entire discussion presupposes that all commodities are traded. For individual commodities it was possible to ask either (i) how net demand or supply responds if one price changes, all other prices remaining constant, or (ii) how net demand or supply responds if all prices in that category (either importables or exportables) change by the same relative amount, with all prices in the remaining category constant. For the aggregate shown by the offer curves, all prices in each category are assumed either to remain constant or change in the same proportion. Questions such as these, which treat all prices as exogenous, are less relevant in a world in which some commodities are not traded, and in which prices are endogenously linked to dis-

turbances created by any exogenous change in the prices of traded goods. For example, suppose all import prices are constant, all export prices rise, and the prices of nontraded goods adjust as required to clear their markets. It is still legitimate to construct offer curves, but what interpretation can be placed on the elasticity of such curves? The breakdown of the ϵ or γ given earlier does not suffice to capture the feedback effect on exports and imports of endogenous changes in the prices of nontradeables.

Without plunging into further formal analysis it is still possible to suggest how crucial is the role of income effects in showing how export supply could be backward bending. But these income effects could be indirect. Ignore aggregation problems by supposing only one exportable and one importable. If the price of exports rises, the quantity of exports could fall if the income effect is strong enough to encourage sufficiently greater local demand for the exportable. Equation (14) sets out the condition. But suppose such direct income effects are ruled out; assume none of the exportable is consumed at home. Does this imply that export supply must rise as export price rises? No, because the price of nontradeables could rise sufficiently to draw resources away from production of exportables (and therefore lower exports).

A simple model to illustrate this was provided by one of the authors of this paper. No importables are produced at home and no exportables are consumed locally. A rise in export price raises real income, tending to raise the price of nontradeables. If this income effect is strong enough, the price of nontradeables could rise even more, relatively, than the export price rise. In such a case the production of exportables falls. The offer curve would show inelasticity of demand for imports and falling export supply (with a rise in export price) because of the way income effects have worked indirectly (through the rise in the price of nontradeables).

One final question can be raised in the context of a model with many exportables and importables. Ignore the presence of nontraded commodities. How does the rise in one import price affect the volume of other imports? Or how does an improvement in one export price affect other export activities? There is an asymmetry in the answer—the same kind of asymmetry underlying the aggregation theorem. If import price i goes up, it is not clear (even in the case of substitutes) how other importables are affected. To some extent resources desert these other sectors and the substitution effect in consumption would tend to raise demand. These effects tending to stimulate other imports are countered by the adverse income effect.

The case of exports is different. A rise in any single export price unambiguously (in the case of substitutes) tends to reduce exports from all other sectors. Local demand for other exportables is stimulated and production curtailed as resources are drawn away. This is the same kind of effect that helped explain how the aggregate export supply elasticity γ could be negative even if all individual γ_i were positive.

REFERENCES

R. W. Jones, "Trade with Non-Traded Goods: The Anatomy of Interconnected Markets," *Economica*, May 1974, *41*, 121–38.

A Note on the Arrow-Lind Theorem

By L. P. FOLDES AND R. REES*

Kenneth Arrow and Robert Lind have recently proved a theorem on risky public projects, stating that under certain conditions the social cost of the risk tends to zero as the population tends to infinity, so that projects can be evaluated on the basis of expected net benefit alone. The present note gives an alternative formulation and a short new proof of the theorem, and uses these to examine the role of certain assumptions concerning the operation of the public sector which in the original were left implicit or received inadequate attention. Some general critical comments on the applicability of the theorem are also offered.

The conditions stated by Arrow and Lind as sufficient for the validity of their result include the following: (i) the government initially appropriates all benefits and pays all costs, distributing the net returns subsequently "through changes in the level of taxes" (p. 371); (ii) the net returns are statistically independent of each person's disposable income in the absence of the project; and (iii) each person's share of the net returns tends to zero as the number n of persons tends to infinity. The result is proved formally only for the case where "all taxpayers [are] identical in that they [have] the same utility function, their incomes [are] represented by identically distributed variables, and they [are] subject to the same tax rates"; but the authors state that ". . . the basic theorem still holds, provided that as n becomes larger the share of the public investment borne by any individual becomes arbitrarily smaller" (p. 373).

This theorem, if generally applicable, would have important practical consequences. It would tend to support an extension of public sector investment by justifying the use of a riskless discount rate applied to expected returns. It would also argue in favor of state participation in private investments where this allows risks to be spread over a larger number of persons. A review of the explicit assumptions alone must cast doubt on the general validity of such applications. The assumption of independence is unrealistic for many investments, for example in infrastructure and "basic" industries whose returns are highly correlated with national income. The assumption that the share of the net benefits of an investment accruing to any person becomes negligible as population tends to infinity is unacceptable in at least three cases: for public goods, where the benefit is not "shared" but increases with the population; for projects whose scale must be adjusted roughly in proportion to the size of population (such as the construction of a grid system of electricity distribution); and for projects whose benefits accrue wholly or in part to a section of the population which is "small" in the sense of the theorem. The last reservation applies not merely to those projects which are specifically designed to benefit only a small part of the population, but also to those special benefits and costs from any project which happen to accrue unavoidably to limited groups. Arrow and Lind avoid this problem in their formal discussion by assuming that the government taxes all benefits and compensates all losses, although they acknowledge that this is unrealistic.

Be that as it may, the present note accepts the Arrow-Lind approach more or less on its own terms, and considers more fully the role of certain implicit assumptions concerning the fiscal system and public expenditure. Specifically, it will be recalled that Arrow and Lind work with only two random variables, the disposable income of a typical individual and the income from distribution of project returns by the government. Although the latter is referred to in-

*London School of Economics, and Queen Mary College, London, respectively. We would like to thank members of the Public Sector Economics Seminar at Queen Mary College for helpful discussion of this paper

formally as representing "changes in the level of taxes," the actual model makes no mention of ordinary taxes or government spending. Project returns are simply appropriated by government, presumably through lump sum compensating taxes, and then distributed to taxpayers by way of a "100 percent dividend," i.e., by lump sum transfers in fixed proportions absorbing the whole of the return. Suppose that this method of distribution were replaced by a more realistic system, for example a variation in the rate of a proportional income tax (gross incomes being regarded as random variables unaffected by the rate of tax). It can be shown that this change would make no essential difference in the Arrow-Lind model *as it stands* because individuals are statistically identical, so that in terms of expected utility each would gain on the swings what he lost on the roundabouts. When this assumption is abandoned, the change will benefit some taxpayers and hurt others. More significantly, a person's disposable income without the project and the effect of the project on that income may become dependent random variables even though the gross income and the total return on the project are independent.[1] Such dependence would, of course, vitiate application of the Arrow-Lind theorem The point at issue here is not only that the repercussions of the project revenue through the system of taxation may create statistical dependence which would not otherwise have existed, but also that the Arrow-Lind assumption of independence owes some of its apparent appeal to the "unnecessary" condition that people are identical The analysis can indeed be extended if this condition is discarded, but the substantive content of the assumption of independence must be considerably strengthened. Further ramifications arise if allowance is made (a) for the possibility of using project returns to finance changes in public expenditure as well as in rates of taxation; (b) for the necessity to treat expenditure or tax rates or both as random variables if the government is to balance its budget; (c) for the possibility that individuals obtain differential benefits from public spending; and (d) for the possibility that individuals derive some direct benefits from projects which are not offset by lump sum taxes, and that these benefits or the public "dividends" or both are subjected to ordinary income taxes.

The model considered below incorporates these features. The general conclusion is that the conditions for the validity of the Arrow-Lind theorem are considerably more stringent than is apparent from the original exposition, and that the circumstances in which the conclusions of the theorem apply are extremely restricted.

I. An Alternative Proof of the Arrow-Lind Theorem

Turning now to a mathematical discussion, we begin with a formulation which follows Arrow and Lind's "implicit" treatment of the public sector but does not assume that persons are identical or that project benefits are initially appropriated by the state. A proof of the main theorem is given which is in some respects simpler and more general than the original one, though it does require that marginal utility be continuous. Structural assumptions about the public sector are then introduced and some of their implications noted.

The economy contains n persons $i = 1, \ldots, n$, where n may take values n_0, $n_0 + 1$, $\ldots$ starting with some n_0. Several random variables will be defined, all of which are supposed to have as domain the same probability space; an elementary event ω corresponds to a state of nature influencing the economy. All random variables are assumed to be integrable, i.e., to have finite expectations. The random variable $x_i = x_i(n)$ represents i's income in the absence of a certain public project when population is n, while $x_i(n) + r_i(n)$ represents his income if the project is introduced. These incomes are defined after all taxes, subsidies, and other effects of government are taken into account, so

[1]It should be noted that the impact of the system of public finance may well be to create negative correlation between project benefits and net disposable incomes, so that by ignoring this influence one may understate the value of public projects to a community of risk averters

that r_i is the total impact of the project upon i's income, including repercussions through the public sector.[2]

One natural interpretation of this setup (though not the only one) is to regard $r_i(n)$ as i's share of the random variable $Z(n)$ representing total returns to the project, whether obtained directly or through changes in taxation; in this case $Z = \Sigma_i r_i$, and all r_i have the same sign as Z. An even more special case would be to assume that all x_i are identically distributed and invariant to n while $r_i(n) = Z(n)/n$ for each i.

Suppose now that the preferences of i among risky incomes are defined by the expected values $Eu_i(.)$ of his utility function, which is assumed to be defined and finite on the real line, strictly increasing, continuously differentiable, and such that $Eu_i(x_i)$ and $Eu_i(x_i + r_i)$ are finite for each n (possibly because u_i is bounded) For given n, i is made better off by the project if and only if,

$$0 \leq E\{u_i(x_i + r_i) - u_i(x_i)\} \tag{1}$$

Applying the mean value theorem for derivatives (separately for each elementary event ω) to the expression under the expectation sign, the condition becomes

$$0 \leq E\{r_i u_i'(x_i + \theta_i r_i)\} \tag{2}$$

where for each ω the value $\theta_i(\omega)$ lies in $[0, 1]$.[3] On multiplying by n, (2) becomes

$$0 \leq E\{nr_i u_i'(x_i + \theta_i r_i)\} \tag{3}$$

Now assume that for a given i as $n \rightarrow \infty$,

(a) there are integrable limiting random variables R_i and $\bar{x}_i$, independent of one another, such that $nr_i(n) \rightarrow R_i$ and $x_i(n) \rightarrow \bar{x}_i$ with probability one; and

(b) passage to the limit under the expectation sign on the right-hand side of (3) is permissible.[4]

For example, (a) obviously holds in the special case mentioned above where $r_i(n) = Z(n)/n$. On the other hand (a) is not satisfied if r_i represents i's untaxed benefit from a pure public good, since then r_i does not vary with n and $|nr_i| \rightarrow \infty$ unless $r_i = 0$.

The assumptions are used as follows. It is inferred from (a) that $r_i(n) \rightarrow 0$ with probability one, hence that $u_i'(x_i + \theta_i r_i) \rightarrow u_i'(\bar{x}_i)$ since u_i' is continuous. Then (a) further shows that the expression under the expectation sign in (3) tends to $R_i u_i'(\bar{x}_i)$, so that by virtue of (b) the expectation itself tends to $E\{R_i u_i'(\bar{x}_i)\}$. This in turn equals $ER_i Eu_i'(\bar{x}_i)$ because R_i and $\bar{x}_i$, hence R_i and $u_i'(\bar{x}_i)$, are independent. Thus, as $n \rightarrow \infty$, the condition (3) for an increase in expected utility becomes

$$0 \leq ER_i Eu_i'(\bar{x}_i) \tag{4}$$

or simply $0 \leq ER_i$ since $u_i' > 0$. To sum up, *when n is "large," i is made better off by the project if, and only if, the expected change in his income is positive.* If we assume with Arrow and Lind that $R_i = Z$ (invariant to n) for every i, then everyone is made better off, if and only if the total expected return from the project is positive Clearly this latter result could be obtained from a weaker assumption, for example that $\lim E Z(n)$ exists and that each $E R_i$ has the same sign as this limit.[5]

[2]In allowing x_i as well as r_i to vary with n, we depart from Arrow and Lind's treatment. This change is necessary because the assumption that income after taxes, etc is invariant to population would impose too many implicit restrictions on the working of the public sector.

[3]Since x_i and r_i are random variables, i.e., measurable functions of events, the same is evidently true of $u_i(x_i + r_i) - u_i(x_i)$ and hence of $r_i u_i'(x_i + \theta_i r_i)$ It can also be shown that $x_i + \theta_i r_i$ and θ_i are random variables, but this fact does not seem to be required

[4]Various conditions can be invoked to justify this operation For example, if u_i is a bounded function of income and the random variables $nr_i(n)$ are uniformly bounded, the result follows from the Dominated Convergence Theorem Weaker assumptions may suffice in particular cases Incidentally, some condition of this kind is needed to justify the passage to the limit in Arrow and Lind's equation (21)

[5]The discussion in the text derives conditions in which a public project increases the welfare of one person Similar methods can clearly be used to obtain conditions for an increase in the value of a social welfare function of the form

$$W = \sum_{i=1}^{n} \alpha_i(n) Eu_i$$

where $\alpha_i(n) > 0$ and $\sum_{i=1}^{n} \alpha_i(n) = 1$ for each n

II. Relevance of the Fiscal System

We now specify the model more explicitly in order to answer some of the questions raised in the introduction; various specifications could be chosen, but the discussion which follows is illustrative of the general results. For each n, let the random variable $G^0(n)$ denote government expenditure in the absence of the project, and $G^1(n)$ the corresponding variable if the project is undertaken. (Distributions of project benefits are excluded from G^1, although the line between these payments and general transfers may have to be drawn arbitrarily.) The value of the benefits (free of tax) which i derives from expenditure $G(n)$ is assumed to have the form $c_i(n)G(n)$, where for each n the c_i are nonnegative constants. If all expenditure is devoted to pure public goods, all constants equal unity; whereas they sum to unity if the government distributes purely private goods. Next, let $X_i(n)$ be i's random income before taxes from private sources; $t^0(n)$ the proportional random rate of tax payable on this income in the absence of the project; and $t^1(n)$ the corresponding rate if the project is undertaken. We write $X(n) = \Sigma_i X_i(n)$ for total gross income from private sources: $a_i(n) = X_i(n)/X(n)$ for i's share of this total. For simplicity we ignore borrowing and lending, and assume that personal and government budgets balance. In the absence of the project, the budget identities are

$$(5) \qquad x_i = (1 - t^0)X_i + c_i G^0$$

$$(6) \qquad G^0 = t^0 X$$

These are identities between random variables, holding for $n = n_0, n_0 + 1, \ldots$. Note that since X is random, G^0 or t^0 or both must be random.

Now let the random variable $Z(n)$ denote total net benefit from the project, $z_i(n)$ the net benefit accruing to i (whether directly or through distributions by government), and $Z_G(n) = Z - \Sigma_i z_i$ the portion retained by the government for the finance of its expenditure. In general the variables z_i need not all have the same sign as Z, but Z_G is assumed to have this sign and not to exceed Z in absolute value. The $z_i(n)$ are supposed to be subject to a random proportional tax rate $\tau(n)$, where $0 \leq \tau(n) \leq 1$. Then the budget identities for persons and government when the project is undertaken have the form

$$(7) \qquad x_i + r_i = (1 - t^1)X_i + c_i G^1 + (1 - \tau)z_i$$

$$(8) \qquad G^1 = t^1 X + \tau(Z - Z_G) + Z_G$$

An expression for r_i can now be obtained from (5) and (7); then (6) and (8) can be used to eliminate either $t^0 - t^1$ or $G^0 - G^1$, yielding respectively,

$$(9) \qquad r_i = (c_i - a_i)(G^1 - G^0) + (1 - \tau)z_i + a_i[\tau Z + (1 - \tau)Z_G]$$

$$(10) \qquad r_i = (c_i - a_i)(t^1 - t^0)X + (1 - \tau)z_i + c_i[\tau Z + (1 - \tau)Z_G]$$

For brevity we now consider two special cases of the model.

A. *Government Expenditure Unaffected by the Project*

In this case $G^1(n) = G^0(n)$ for each n, and so (9) reduces to

$$(11) \qquad r_i = (1 - \tau)z_i + a_i[\tau Z + (1 - \tau)Z_G]$$

while (5) and (6) together with $a_i = X_i/X$ yield

$$(12) \qquad x_i = a_i(X - G_0) + c_i G^0 = (na_i)(X - G^0)/n + (nc_i)(G^0/n)$$

In order to apply condition (a) of our proof of the Arrow-Lind Theorem to the values of nr_i and x_i appearing in (11) and (12), it is necessary to make assumptions about the limits of several variables as $n \rightarrow \infty$.

First, to ensure that $nr_i(n)$ converges to some R_i, it is enough by (11) to assume that nz_i, na_i, τ, Z, and Z_G converge (with probability one, to integrable limiting variables). Note that the convergence of $na_i = nX_i/X$ implies that the income share a_i of a given individual tends to zero; this holds, for example, in the special case where $X_i = X/n$. Note also that the case of pure public goods is ruled out by the assumptions about the project variables Z and nz_i.

Secondly, to ensure that $x_i(n)$ converges to some $\bar{x}_i$, it is enough by (12) to assume that

X/n, G^0/n, and nc_i (as well as na_i) converge. In other words, there must be limits with finite expectations of private sector income and government expenditure per head, and of the product of population with i's benefit per dollar of public expenditure; the first two conditions seem reasonable, while the last may be subject to reservations if the level of spending on public goods is maintained when the population is large.

Thirdly, the limiting variables R_i and $\bar{x}_i$ have to be independent; this presents difficulties since by (11) and (12) the term na_i is common to nr_i and x_i. To rule out dependence, it is enough to adopt one of the following assumptions:

(i) That lim na_i is degenerate (constant across states); this appears too special a condition to be acceptable.

(ii) That $\lim (X - G^0)/n = 0$; this would be satisfied if $G^0/X \rightarrow 1$ with X/n bounded, or if $X/n \rightarrow 0$. The former case implies a tax rate of $t^0 = G^0/X$ rising to 100 percent as population increases, which is unreasonable. The latter case, that gross private income per head tends to zero, could well occur for Malthusian reasons; even so, applications of the theorem must be made for finite populations, and it is unlikely that $(X - G^0)/n$ would in practice be close enough to zero.

(iii) That $\lim [\tau Z + (1 - \tau Z_G] = 0$; this holds in general only if $\lim \tau = \lim Z_G = 0$ (leaving aside the uninteresting case in which $\lim Z = 0$). This means that, in the limit, no tax is imposed on project benefits, and no part of them is retained by the government to finance expenditure; in other words *the benefits of the project accrue in full to persons, without liability to tax.*

The only one of these assumptions which has real economic interest is that in (iii). If we adopt it, we are accepting that the public project in question has no fiscal repercussions, a condition which in many practical cases would not be satisfied. For example, an irrigation project would generate additional incomes for farmers, and these would be subject to tax. Again, the returns from investment in British nationalized industries are viewed as inflows to the public sector, and are regarded for many purposes as a source of finance in much the same way as indirect taxation. Suppose that we accept the assumption nevertheless; then (11) reduces to $r_i = nz_i$, and it remains to postulate that $R_i = \lim nz_i$ is independent of the limits of $(X - G^0)/n$, c_iG^0, and $na_i = nX_i/X$. Thus, in the limit, *project income for each person to whom the theorem is applied is to be independent not only of the per capita difference between gross private sector income and public expenditure, but also of the person's benefit per unit of public expenditure and of the ratio of private sector income to the population average.* While there seems to be no theoretical relationship among the variables which rules out such independence a priori, the stated condition is clearly much more restrictive than the simple requirement that the project return be independent of each person's gross income from private sources.

B. *Tax Rates Unaffected by the Project*

In this case $t^1 = t^0$, so that

$$r_i = (1 - \tau)z_i + c_i[\tau Z + (1 - \tau)Z_G] \tag{13}$$

from (10), while (5) and (6) yield

$$x_i = (na_i)(X - t^0X)/n + (nc_i)(t^0X/n) \tag{14}$$

The analysis now proceeds along much the same lines as under Section IIA, except that government expenditure per head G^0/n is replaced by tax revenue per head t^0X/n. Briefly, to ensure the convergence of nr_i and x_i we assume integrable limits for nz_i, na_i, τ, Z, Z_G, X/n, nc_i, and t^0X/n. As regards the independence of the limits, it is now nc_i which appears in both expressions, and we are led as before to the unattractive assumption that $\lim \tau = \lim Z_G = 0$. It then remains to suppose that $R_i = \lim nz_i$ is independent of the limits of $(X - t^0X)/n$, c_it^0X, and na_i; thus *public expenditure per head is replaced by tax revenue per head in the independence condition given above.*

A striking feature of the various italicized conditions in this section is that the aggregative assumption of independence adopted by Arrow and Lind can reasonably be expected to hold only if similar assumptions are satisfied by each of a number of sectoral variables. These assumptions must, of course, hold separately for each individual to whom the theorem is applied. This not only represents a strengthening of assumptions, but also makes it much more difficult to establish that the conditions under which the theorem holds are met in any given situation.

III. Conclusions

Using a proof of the Arrow-Lind theorem which makes the roles of the various assumptions more transparent, we have tried in this note to bring out the implications of a more realistic specification of the fiscal system in which public sector investment is embedded. Taxes and expenditure exist with or without any one project, and the government must balance its budget in each state of the world. We find that in this case the sufficient conditions used in the proof of the theorem become a good deal more restrictive. In particular, to ensure independence between the impact of a project on the individual's income and his marginal utility across states of the world we have to assume—leaving aside some unappealing special cases—that the project income is free from taxation and that none of it is retained to finance public expenditure. The assumption of independence between the project's return and private incomes, which itself is open to question, must be extended to a number of sectoral variables in a way for which there is no obvious empirical justification. These results suggest that there is still a need for an analysis of public sector investment criteria under uncertainty,[6] which will yield more general and robust results.

[6]The papers by Rees (1973, 1976) adopt more general approaches to this problem

REFERENCES

K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

R. Rees, "Public Sector Resource Allocation Under Risk," in Michael Parkin, ed., *A.U.T.E Essays in Modern Economics*, London 1973.

———, "Uncertainty, Second Best and Public Policy," disc. pap. no. 28, econ. dept., mimeo, Queen Mary College 1976.

On Returns to Scale and the Stability of Competitive Equilibrium

By GÉRARD GAUDET*

Much empirical evidence has been obtained on returns to scale by estimating relationships derived from the first-order conditions for profit maximization under assumptions of perfect competition.[1] Results showing estimates of scale parameters significantly greater than unity have usually been criticized or simply rejected on the grounds that with increasing returns to scale the first-order marginal conditions are inappropriate since the corresponding second-order conditions are not satisfied.

The purpose of this note is to show that under quite plausible conditions, production in the increasing returns range of the production function may indeed be compatible with a stable competitive equilibrium. The conditions postulated here are that of a firm operating in a dynamic context and facing costs explicitly associated with the adjustment of its stock of capital, costs which do not necessarily vanish in long-run equilibrium.[2]

Assume that the cost to the firm of adjusting its stock of capital is not independent of the speed of adjustment. I treat this premium to speed as an addition to the firm's costs. Thus investment expenditures at the rate of gross investment I are $qI + pC(I)$, where $pC(I)$ represents adjustment costs, q is the supply price of investment goods, and p is output price.[3] For simplicity, a Cobb-Douglas production function will also be assumed, i.e.,

$$(1) \qquad F(K, L) = \gamma K^{\alpha} L^{\beta}$$

where $0 < \alpha, \beta < 1$, but the scale coefficient (or elasticity of scale) $\epsilon = \alpha + \beta$, can take on any positive value.

The firm is assumed to be a present-value maximizer. It will maximize the sum of discounted future net cash flows,

$$(2) \quad PV = \int_0^{\infty} e^{-rt} \{p\gamma K^{\alpha} L^{\beta} - wL - qI - pC(I)\}\, dt$$

subject to the constraints

$$(3) \qquad \dot{K} = I = \delta K \qquad K(0) = K_0$$

The Euler-Lagrange equations for this optimization problem can be written

$$(4a) \qquad \dot{K} = I - \delta K$$

$$(4b) \quad \dot{I} = \frac{(r + \delta)[q + pC'(I)] - p\, a\, K^{(1-\epsilon)/(\beta-1)}}{pC''(I)}$$

*Assistant professor, department of economics, Université Laval, Québec. The preparation of the final draft of the paper has benefited from the comments of an anonymous referee

[1] For a recent survey, see Dale W. Jorgenson.

[2] The theory of the firm with adjustment costs is developed in Robert E. Lucas, Arthur Treadway (1969, 1970), and J. P. Gould, among others.

[3] The most general form in which adjustment costs of the type assumed here have been introduced is by writing the production function in the form $G(K, L, I)$, (see for example Lucas, p. 325) I have adopted the separable form $G(K, L, I) = F(K, L) - C(I)$, mainly because it is much more manageable for the problem at hand. The form assumed here in fact resembles that used by Gould except that he embeds the term qI into a more general $C(I)$ function I chose not to do so here in order to make explicit that marginal investment cost is not zero at $I = 0$, a fact which may then be used to eliminate some possibilities in drawing the illustrative phase diagram. The reader will observe however that these assumptions are purely simplifying in nature and do not in any way change the main conclusion of the paper For a further discussion of the various functional forms under which adjustment costs have been introduced in the theory of the firm in the literature, see Treadway (1970)

where L has been solved as a function of K and the real wage from the marginal productivity condition for labor, and where

$$a = \gamma\alpha \frac{w}{\gamma\beta p}\beta/(\beta - 1) \tag{5}$$

The first and second derivatives of $C(I)$ are denoted by $C'(I)$ ($\gtreqless 0$ as $I \gtreqless 0$) and $C''(I) > 0$, respectively.

Now let $(\bar{K}, \bar{I})$ denote an arbitrary stationary of system (4). Then

$$\bar{I} - \delta\bar{K} = 0 \tag{6a}$$

$$(r + \delta)[q + pC'(\bar{I})] - p\,a\,\bar{K}^{(1-\epsilon)/(\beta-1)} = 0 \tag{6b}$$

Linearizing system (4) around $(\bar{K}, \bar{I})$ we find that its characteristic roots are given by

$$\theta = r/2 \pm \left[(r/2)^2 + (r + \delta)\delta - \frac{a(1 - \epsilon)\bar{K}^b}{(\beta - 1)C''(\bar{I})}\right]^{\frac{1}{2}} \tag{7}$$

where

$$b = [2(1 - \beta) - \alpha]/(\beta - 1) \tag{8}$$

If

$$(r + \delta)\delta > \frac{a(1 - \epsilon)\bar{K}^b}{(\beta - 1)C''(\bar{I})} \tag{9}$$

then both roots are real and of opposite signs, and therefore $(\bar{K}, \bar{I})$ is a saddle point. Otherwise the roots are either real and both positive or complex with positive real parts, and $(\bar{K}, \bar{I})$ is locally unstable. We notice first that if $\epsilon < 1$ or $\epsilon = 1$ inequality (9) is satisfied, and it follows that $(\bar{K}, \bar{I})$ is necessarily a saddle point, and the optimal investment path is the stable arm of the saddle point. But notice also that inequality (9) imposes no restriction on the sign of $(1 - \epsilon)$: we may well have a stable equilibrium (i.e., saddle point) with $\epsilon > 1$. Although an unstable equilibrium can occur only with increasing returns to scale, the existence of increasing returns ($\epsilon > 1$) does not rule out the possibility of a stable equilibrium.

The case where $\epsilon < 1$ is illustrated in phase space in Figure 1. In that case it is easily verified from equation (6b) that the slope of the $\dot{I} = 0$ locus is

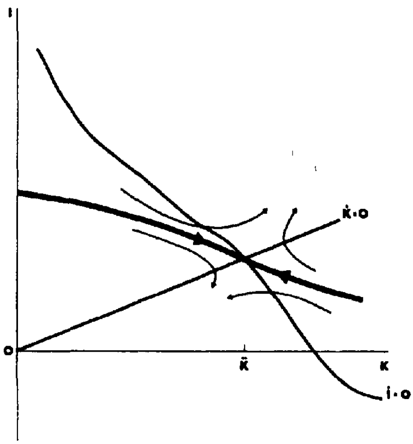


FIGURE 1

$$\frac{dI}{dK} = \frac{a(1 - \epsilon)K^b}{(\beta - 1)(r + \delta)C''(I)} < 0 \tag{10}$$

The slope of the $\dot{K} = 0$ locus is of course always equal to $\delta > 0$. In that case there exists a unique stationary which is always a saddle point, as is verified from inequality (9). The same can be said of the case where $\epsilon = 1$, in which case the slope of the $\dot{I} = 0$ locus is zero.

Things are different with $\epsilon > 1$. In that case the slope of the $\dot{I} = 0$ locus is positive and, with the Cobb-Douglas production function, if there exists only one stationary it is an unstable one since the $\dot{I} = 0$ locus would necessarily cut the $\dot{K} = 0$ locus from below at this unique stationary.[4] This follows from equation (6b) from which we verify that

[4] With a more general form for the production function the second term of equation (6b) will not necessarily be zero at $K = 0$, thus allowing for the possibility of a stable unique stationary in the increasing returns range of the production function. Such a case is illustrated in my cited research paper.

$$(11) \qquad \lim_{K \to 0} C'(I) = -q/p$$

which implies

$$\lim_{K \to 0} \dot{I} < 0$$

along the $\dot{I} = 0$ locus. However more than one stationary may exist with $\epsilon > 1$, in which case every second stationary is a saddle point since the $\dot{I} = 0$ locus will have to cut the $\dot{K} = 0$ locus from above at those points. Such a possibility is illustrated in Figure 2, where $(\bar{K}_1, \bar{I}_1)$ and $(\bar{K}_2, \bar{I}_2)$ are respectively locally unstable and stable, and the heavy arrow indicates the optimal path. The case drawn assumes the roots to be real at $(\bar{K}_1, \bar{I}_1)$. Another possibility would be for the roots at $(\bar{K}_1, \bar{I}_1)$ to be complex with positive real parts, resulting in an unstable spiral around $(\bar{K}_1, \bar{I}_1)$, but our point is simply that a stable stationary (i.e., a saddle point) is not inconsistent with $\epsilon > 1$.

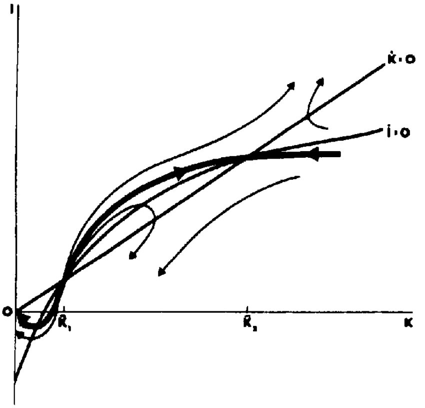


FIGURE 2

This fact may be further illustrated by considering an equivalent static profit-maximization problem. Thus assume that the firm is seeking the values of L and K that will maximize profits every period. Once this optimal K is attained gross investment will equal replacement investment, i.e., $I = \delta K$.

Now the cost to the firm of a unit of investment good, at the rate of investment I, can be written $q + pC(I)/I$ which becomes $q + pC(\delta K)/\delta K$ when $I = \delta K$. The profit function can then be written

$$(12) \qquad \pi(K, L) = p\gamma K^{\alpha} L^{\beta} - wL - (r + \delta)[q + pC(\delta K)/\delta K]K$$

which is to be maximized with respect to K and L. Carrying out this maximization we find that the first-order conditions are simply the conditions for stationarity of system (4). As for the second-order (sufficiency) conditions, we easily verify that the Hessian of the profit function is negative definite if and only if inequality (9) is satisfied. This, again, imposes no restrictions on the sign of $(1 - \epsilon)$: the sufficiency conditions may well be satisfied and the equilibrium be stable even with $\epsilon > 1$.

What is brought out by this example[5] is that dynamic considerations may modify the usual relation between the purely technological information contained in the production function and the economic information contained in the cost function. The conclusion as to the instability of a competitive equilibrium with increasing returns is derived from the instability of a competitive equilibrium with decreasing costs, via the one-to-one relationship which can usually be established between decreasing costs and increasing returns. But this last relationship may break up, as is the case in the present example, if one properly accounts for all of the costs which may be attributed to the employment of the factors of production appearing in the production function. The firm may then find itself facing increasing costs of production while still, technically, in the increasing returns range of its production function. This suggests that the proper empirical measure of "returns to scale" may be the shape of the cost function rather than that of the production function.

[5]Since the point is simply that a stable equilibrium is not necessarily incompatible with increasing returns in the production function, the use of the Cobb-Douglas form for the production function suffices in establishing the proof. However similar results can be shown using a more general quasi-concave production function, the proof of which is sketched in the Appendix

APPENDIX

If we replace the Cobb-Douglas production function by a more general quasi-concave production function, which we will assume twice continuously differentiable with positive and diminishing marginal productivities with respect to both factors, then upon linearizing the equilibrium conditions around an arbitrary stationary $(\bar{K}, \bar{l})$ we find that the characteristic roots are given by

$$\text{(A1)} \qquad \theta = \frac{r}{2} \pm \left[\left(\frac{r}{2}\right)^2 + (r+\delta)\delta - \frac{|H(\bar{K},\bar{L})|}{C''(\bar{l})F_{LL}(\bar{K},\bar{L})}\right]^{\frac{1}{2}}$$

where $H(K, L)$ is the Hessian matrix of the production function and

$$\text{(A2)} \qquad \bar{L} = L(\bar{K}, w/p)$$

from the marginal productivity condition for labor.

Again if and only if

$$\text{(A3)} \qquad (r+\delta)\delta > \frac{|H(\bar{K},\bar{L})|}{C''(\bar{l})F_{LL}(\bar{K},\bar{L})}$$

the roots are real and of opposite signs and $(\bar{K}, \bar{l})$ is a saddle point. The denominator of the right-hand side being negative then clearly if $|H(\bar{K},\bar{L})|$ is positive, inequality (A3) is satisfied. However inequality (A3) does not impose a sign on $|H(\bar{K},\bar{L})|$ and may well be satisfied with $|H(\bar{K},\bar{L})|$ negative. Thus stability does not require concavity of the production function in the neighborhood of equilibrium.

There remains to associate a negative Hessian determinant to the increasing returns range of the production function. The usual definition of the elasticity of scale ϵ of the production function is

$$\text{(A4)} \qquad \epsilon = \left.\frac{\partial \log F(kK, kL)}{\partial \log k}\right|_{k=1} = \frac{F_K K + F_L L}{Q}$$

where

$$\text{(A5)} \qquad Q = F(K, L)$$

Define an "isocline" (see John Rowe) as the locus of points (K, L) satisfying (A5) as well as

$$\text{(A6)} \qquad \lambda F_i(K, L) = v_i \qquad i = K, L$$

where v_i is an arbitrary positive constant. If factor prices were constant and represented by v_i then the isocline would constitute an expansion path. We can show (see Giora Hanoch, p. 494) that along such an isocline

$$\text{(A7)} \qquad \frac{\partial \epsilon}{\partial Q} = \left(\frac{1}{Q}\right)(1-\epsilon) - \epsilon\frac{|H(K,L)|}{|B(K,L)|}$$

where

$$B = \begin{bmatrix} 0 & F_K & F_L \\ F_K & F_{KK} & F_{KL} \\ F_L & F_{KL} & F_{LL} \end{bmatrix}$$

and $|B| > 0$ by quasi concavity. Now if we assume, as is typical, that the production function exhibits increasing followed by constant and decreasing returns and that $\partial\epsilon/\partial Q \leqq 0$ along an isocline, then from (A7) we verify that $|H| < 0$ implies $\epsilon > 1$.

Note that for a homogeneous production function ϵ is equal to the degree of homogeneity of the function, the isocline is a ray through the origin, and $\partial\epsilon/\partial Q = 0$. Thus it follows directly in that case that ϵ is greater or smaller than one as $|H|$ is negative or positive.

An alternative approach to relating returns to scale to the sign of $|H|$ is to define, following Rowe, "marginal returns to scale" with respect to the spacing of isoquants. It can then be shown (see Rowe, p. 550) that isoquants become closer, are evenly spaced, or become farther apart along an isocline, and thus marginal returns to scale are increasing, constant, or decreasing as $|H|$ is negative, zero, or positive.

REFERENCES

G. O. Gaudet, "Adjustment Costs and Optimal Firm Size," res. rept. 7309, dept. econ., Univ. Western Ontario 1973.

J. P. Gould, "Adjustment Costs in the Theory of Investment of the Firm," *Rev. Econ. Stud.*, Jan. 1968, *35*, 47–56.

G. Hanoch, "The Elasticity of Scale and the Shape of Average Costs," *Amer. Econ. Rev.*, June 1975, *65*, 492–97.

D. W. Jorgenson, "Investment Behavior and the Production Function," *Bell J. Econ.*, Spring 1972, *3*, 220–51.

R. E. Lucas, "Adjustment Costs and the Theory of Supply," *J. Polit. Econ.*, Aug. 1967, *75*, 321–34.

J. Rowe, "Returns to Scale and the Spacing of Isoquants: Comment," *Amer. Econ. Rev.*, June 1968, *58*, 548–50.

A. B. Treadway, "On Rational Entrepreneurial Behavior and the Demand for Investment," *Rev. Econ. Stud.*, Apr. 1969, *36*, 227–39.

———, "Adjustment Costs and Variable Inputs in the Theory of the Competitive Firm," *J. Econ. Theory*, Dec. 1970, *2*, 329–47.

The Earnings and Promotion of Women Faculty: Comment

By STEPHEN FARBER*

Several studies have investigated salary and academic rank differentials among academic professionals using cross-sectional surveys. In spite of the variation in approaches to the issue of sex differences in salary and academic rank, none of the studies have utilized longitudinal surveys. The purpose of this study is to test whether the results of cross-sectional studies of sex differences in academic rank promotion and earnings profiles are similar to results derived from longitudinal data.

In a cross-sectional study of doctorate scientists employed in universities and colleges in 1964, Allan Bayer and Helen Astin compared sex differences in salary and academic rank for two experience cohorts—one with 2 years of postdoctoral experience, the other with 6 years of postdoctoral experience. They found that ". . . women in the natural sciences emerge as receiving promotions on par with men" (p. 199). They also found that women receive initially lower salaries than men, but that "no consistent increases or decreases in sex differences in salary emerge over time" (p. 198). Bayer and Astin's study cannot be expected to generate any information about sex differences in salary and rank promotion for individuals with more than 6 years of experience.

In a cross-sectional study of faculty at a large university in 1970, Nancy Gordon, Thomas Morton, and Ina Braden, after controlling for other variables, concluded that women's salaries peaked at an earlier age than those of men. Women's salaries rose more rapidly with age than those of men for ages less than 35. Between 35 and 40 women's salaries were flat, and declined after age 40 (p. 423). They found that the longer the faculty member was employed by the university, the lower was the relative salary of the female.

In the most extensive study of sex differences in academic salaries, George Johnson and Frank Stafford, using a national sample of college faculty, apply the well-known human capital model of the earnings function to estimate the cross-sectional effects of sex on the level and shape of the experience-earnings stream. They suggest that a life cycle difference in labor force participation exists between male and female doctorates: females reduce their full-time labor force participation for child-rearing then reenter the labor force as their children reach school age. They cite evidence which supports this hypothesis (p. 891). They also note that females work fewer hours per year than males when they do work (p. 892). The effect is that females will have accumulated fewer years of work experience after receiving their doctorates than males with equal potential work experience. Consequently, the salary differential between males and females is expected to widen with potential experience up to the end of the childbearing years, at which time permanent return to the labor force may stimulate renewed job related investment and cause the salary differential to decline (p. 890). Johnson and Stafford find that the salary differential is smallest immediately after receiving the doctorate, rises until 15 years after receipt of the doctorate, and then "there is typically a narrowing of the female-male salary differential at advanced years of postdegree experience" (p. 895). However, the empirical support for the latter conclusion is weak. In only two of the six fields they studied is the appropriate interaction term significant. An additional explanation of the initial flatter salary profile of females is that females choose employment in schools which do not emphasize

*Assistant professor, Louisiana State University

research since their impermanent attachment to the labor force reduces the incentive to undertake costly training options (p. 897).

The analysis of rank in Johnson and Stafford's study is less complete than their analysis of salary. They presume that rank is simply a more complete index of total compensation than salary; therefore, the life cycle influences which affect salary will similarly affect rank (p. 896). Among biologists, the cumulative probability of achieving a high rank rises more rapidly for males than females during the period from 7 to 19 years after the doctorate. They also cite evidence from another study which showed that more never-married than ever-married females had the rank of full professor 20 years after receipt of the doctorate (p. 896).

I. The Longitudinal Earnings Function

Jacob Mincer and Solomon Polachek's (p. 586) general form of the wage equation is:

$$(1)\quad ln\, E_t = ln\, E_0 + rs + r\sum_{i=1}^{n}\int_{t_i}^{t_{i+1}}(a_i + b_i t)dt$$

where $ln\, E_t$ is the natural log of earnings capacity at time t; $ln\, E_0$ is the natural log of earnings capacity prior to investment in formal schooling and on-the-job training; r is the average rate of return on human capital; s is years of formal schooling, n is the number of separate segments of participation and nonparticipation in the labor force; $a_i + b_i t$ represents the proportion of earnings capacity invested in human capital after t years of the ith segment have elapsed. If the first observation of earnings capacity occurs at time m, after completion of formal schooling, and the investment function is:

$$(2)\quad a_i + b_i t = a + bx \qquad (0 < a < 1,\ b < 0)$$

where x is the number of years of total work experience accumulated by time t of the ith work segment,[1] the longitudinal earnings function has the following form:

$$(3)\quad ln\,\frac{E_{m+1}}{E_m} = r\int_{x_m}^{x_{m+1}}(a + bt)dt$$

where x_m is years of experience at the initial point of observation and x_{m+1} is years of work experience accumulated by the second point in time. Integrating equation (3) yields:

$$(4)\quad ln\,\frac{E_{m+1}}{E_m} = ra(x_{m+1} - x_m) + \frac{rb}{2}(x^2_{m+1} - x^2_m)$$

which represents an empirically estimable form of the longitudinal earnings function. Earnings will increase at a more rapid rate the larger r or a and the smaller the negative value of b. Also, given a constant difference in experience between two points in time, the negative coefficient of the second term implies a less rapid rate of increase in earnings the greater the years of experience x_m.

Several factors which are correlated with sex and should affect the earnings stream must be controlled for. First, Johnson and Stafford (pp. 892–95) found, on average, that females have more predoctoral experience than males. They have hypothesized that individuals who begin postdoctoral work experience later in their life cycle tend to undertake less postdoctoral training and possess flatter earnings streams the longer they delay postdoctoral employment. However, an individual with poor abilities may take many years to complete the doctorate, while another individual with high abilities may undertake considerable work activity prior to completing or entering a doctoral program. The latter individual may find that predoctoral work experience and formal education are complementary investments which increase the rate of return from postdoctoral investment in on-the-job training. Holding constant the age at which the Ph.D. was attained, greater predoctoral experience may be associated with a precocious individual and large self-financed postdoctoral training investment, resulting in a steep earnings profile.

[1] This form of the investment function assumes that human capital investment arises only from job experience after completion of formal schooling and that the level and rate of decline of the investment proportion is independent of nonparticipation in the labor force

Second, Johnson and Stafford's life cycle hypothesis suggests that there may be major sex differences in types of work activity of males and females. Since employers expect women to drop out of the labor force during childbearing stages of the life cycle, they may not be willing to invest in firm-specific work skills for females. Consequently, women may have less access to administrative work activity than males, particularly at younger ages. Similarly, as Johnson and Stafford (p. 897) suggest, females may seek nonresearch positions which do not require large self-financed training costs. Since training for administrative tasks is typically more job specific than teaching and research, individuals engaged in administrative work activity are expected to have flatter earnings streams than teachers and researchers.

Third, human capital investment is not always in the form of work experience. Another important form of investment is job mobility. A voluntary change of employers and a voluntary change from one type of work activity to another represent human capital investments. If females have a lower incentive to invest in themselves than males, they may undertake less voluntary migration and occupational change. These two forms of investment are expected to steepen the earnings profile.

II. Sex Differences in Salary Increases

This study uses a National Science Foundation *(NSF)* panel data set of approximately 12,000 doctorate scientists employed in a college or university in 1960. These scientists are a subset of the *NSF* longitudinal doctorate records file. Individuals in this file responded to *NSF* biennial surveys in 1960, 1962, 1964, and 1966. They are in the following fields: agriculture, biology, chemistry, earth sciences, mathematics, physics, and psychology.

The symbolic form of the basic regression equation is:

$$(5)\quad Y = \alpha_1 + \alpha_2 DX + \alpha_3 DXSQ + \alpha_4 SEX + \alpha_5 SEX \cdot DX + e$$

where the dependent variable is the natural *log* of the ratio of 1966 and 1960 9-month salary;[2] *SEX* is a dummy variable (1- female, 0- male). Only those individuals reporting full-time academic employment in both the *NSF* censuses of 1960 and 1966 were included in this sample. If the individual reported full-time employment during both of the intervening biennial census years 1962 and 1964, it was assumed that $DX = 6$; if the individual was not fully employed in one of the intervening censuses, $DX = 4$; and if the individual was not fully employed during both intervening censuses, $DX = 2$. This measurement of experience controls partially for the possibility that females may be more part-time participants in the labor force than males. There was no distinction made between various types of less than full-time employment. Also, assuming two years of labor force nonparticipation for each year of reported non-full-time experience is somewhat arbitrary. It was impossible to distinguish individuals who were not fully employed in off-census years but were fully employed in the census year. I have also introduced a variable which measures interaction between experience before and after 1960. Accordingly, $DXSQ = 2x_m \cdot DX + DX^2$ which varies with both x_m, years of experience prior to 1960, and *DX*.

The data were divided into three subsets: age in 1960 less than 40 (−40), age in 1960 greater than or equal to 40 and less than 50 (40–50), and age in 1960 greater than or equal to 50 (50+). Since the average age for obtaining the doctorate was 30 in this sample, these age groups intuitively correspond with the childbearing, childrearing, and childfree years of the life cycle. Regressions were run on all age groups combined and on each of the separate groups in Table 1. All regressions show the expected positive sign for *DX* and the negative sign for *DXSQ*.

Before discussing the implications of these regressions for sex difference, it is necessary to point out several interesting results. First, the

[2] In the *NSF* questionnaire, salary in 1960 was not distinguished on a 9- or 12-month basis. However, in 1966 the majority of salary responses were on a 9-month basis.

TABLE 1—REGRESSIONS EXPLAINING THE NATURAL *Log* OF THE RATIO OF 1966 SALARY TO 1960 SALARY BY AGE COHORT

Variable	Total	−40	40–50	50+
Intercept	0.477	0.561	0.455	0.181
	(18.1)[a]	(12.1)[a]	(8.7)[a]	(3.8)[a]
DX	0.033	0.025	0.033	0.029
	(9.4)[a]	(4.6)[a]	(4.5)[a]	(3.1)[a]
DXSQ	−0.001	−0.0008	−0.0007	−0.0003
	(−30.4)[a]	(−8.2)[a]	(−6.4)[a]	(−4.0)[a]
SEX	−0.273	−0.448	−0.165	0.106
	(−3.0)[a]	(−3.2)[a]	(−0.9)	(0.7)
SEX · *DX*	0.047	0.075	0.031	−0.019
	(3.0)[a]	(3.1)[a]	(1.0)	(−0.7)
AGEPHD	−0.066	−0.008	−0.006	0.001
	(−10.2)[a]	(−5.8)[a]	(−4.0)[a]	(3.6)[a]
XPR	0.008	0.007	0.009	0.003
	(11.2)[a]	(4.4)[a]	(5.3)[a]	(3.3)[a]
RES	0.086	0.090	0.091	0.084
	(11.4)[a]	(7.4)[a]	(8.1)[a]	(5.0)[a]
TCH	0.127	0.138	0.128	0.102
	(17.8)[a]	(11.3)[a]	(12.2)[a]	(7.5)[a]
MOBIL	0.010	0.024	−0.013	−0.042
	(1.7)[c]	(3.1)[a]	(−1.2)	(−2.2)[b]
CWA	−0.003	−0.007	0.022	−0.030
	(−0.6)	(−1.0)	(2.8)[a]	(−2.6)[a]
R^2	.11	.04	.06	.08
F	149.9[a]	27.3[a]	21.0[a]	16.0[a]
N	12,163	6681	3595	1951
Number female	578	211	187	180

[a]significant at the .01 level
[b]significant at the .05 level
[c]significant at the .10 level

older the age at which the individual received the doctorate, *AGEPHD*, the lower the rate of salary increase within the youngest two age cohorts. However, the greater the number of years of predoctoral experience, *XPR*, the greater the rate of salary increase. This was true for all age cohorts.[3] This result is contrary to Johnson and Stafford's (p. 894) result that more predoctoral experience is associated with a flatter earnings stream. The effects of job mobility on earnings vary with age. Young individuals who changed their location of employment between 1960 and 1966, *MOBIL* = 1, received significantly higher salary increases than those who did not migrate. The opposite was true for individuals in the oldest age cohort. The middle-age individuals received positive salary benefits from changing their primary work activity between 1960 and 1966, *CWA* = 1; but the opposite was true for the oldest age cohort. These results suggest that older individuals, who may be satisfied with their earnings, become more interested in psychic benefits of the job, such as location or the type of work activity, than pecuniary benefits. These regressions support the hypothesis that individuals engaged primarily in nonadministrative activities, such as research (*RES* = 1) or teaching (*TCH* = 1) will receive higher salary increases than individuals engaged primarily in administrative activities.[4]

After controlling for the above factors, the effect of sex on the earnings stream varies with

[3]*XPR* is calculated the same as Johnson and Stafford's study (p. 892): $XPR = X - XPO$ where X = years of reported experience in 1960; $XPO = 1960 - AGEPHD$. Although *XPR* and *AGEPHD* are correlated in the sample, the t-values of the coefficients remain highly significant in Table 1.

[4]The *NSF* questionnaire asked individuals to report as their primary activity the activity in which they spent more than half of their time.

age. It is only in the youngest cohort that females ($SEX = 1$) received significantly lower rates of salary increases than males. This age coincides with the hypothesized childbearing years. The coefficients for the interaction term, $SEX \cdot DX$, suggest that for younger age females an additional year of full employment between 1960 and 1966 resulted in a greater salary increase than for comparable males. In fact, the coefficient for the interaction term suggests that 6 years of full employment will exactly offset the young female's initially lower salary increase. This implies that a female will receive the same salary increase as a male in the younger age cohort only if she shows that she is a consistent labor force participant. Since females have lower salaries than comparable males (Johnson and Stafford, p. 894), the equal rates of salary increases among the middle- and old-age cohorts mean lower absolute salary increases for females than males within these cohorts.[5]

IV. Sex Differences in Rank Promotion

Johnson and Stafford have suggested that academic rank is an index of total compensation, including salary, fringe benefits, and recognition (p. 896). If this is true, the human capital investment model which generated the longitudinal earnings function, equation (5), should generate an analogous longitudinal function for academic rank promotion between 1960 and 1966. However, the factors which affect salary increases may not have a similar effect on rank promotions. First, a university may be more willing to give a salary increase than promotion to a tenured rank since a tenured individual becomes a fixed cost to the university. Second, individuals may view rank and earnings as substitutable, or rank may be insignificant in making investment decisions.

The regression model which was used to explain salary increases between 1960 and 1966 in Table 1 was used to explain rank promotions over this same period. Two estimation techniques were employed. Both techniques employed a dummy dependent variable which was assigned the value 1 if the individual had a higher academic rank in 1966 than in 1960 and 0 otherwise. The regressions in Tables 2 and 3 show the results of using ordinary least squares (*OLS*). First, regressions were run on the total sample, regardless of age or rank, and on the three age categories. These results are shown in Table 2. With one exception, the regression on the total sample in Table 2 suggests that the factors which determine salary increases determine rank promotions in the same way. The negative sign of the coefficient for *MOBIL* in the total regression in Table 2 suggests that individuals had a lower chance of obtaining a rank promotion by changing academic employers than if they remained immobile. However, when the total sample is broken down by age category, it is apparent that this negative coefficient is due to the negative impact of migration on rank promotion among younger individuals. This is consistent with the fact that the likelihood of involuntary moves is greater among younger, lower ranked professors (see David Brown, p. 43). The insignificance of the coefficients for the geographic mobility variable in the two older age categories is misleading. A high correlation between geographic mobility and change in work activity, *CWA*, makes it difficult to determine the effect of either of these two forms of human capital investment on rank promotions.

The effect of sex on rank promotions is the same as the effect of sex on salary increases for the total sample. When the sample is broken down into the three age cohorts, females received significantly fewer rank promotions in the younger age cohort.[6] This was also true of salary increases. However, for the middle-age cohort, females received significantly more rank pro-

[5]It would have been interesting and useful to perform separate regressions of the form shown in Table 1 for each of the seven disciplines. However, disaggregation by discipline resulted in too few observations of females in each discipline to be meaningful. Since some individuals did not report their discipline, introducing a discipline dummy variable would result in a reduction in total sample size from 12,163 to 8,752. A Chow test performed on the 8,752 observations showed a significant structural difference in the regression coefficients across disciplines.

[6]Because of the collinearity between SEX and $SEX \cdot DX$ in the middle-age cohort, $SEX \cdot DX$ is excluded from the regression of the middle-age cohort.

TABLE 2—*OLS* REGRESSIONS EXPLAINING THE DICHOTOMOUS VARIABLE RANK PROMOTION BETWEEN 1960 AND 1966 BY AGE COHORT

Variable	Total	−40	40–50	50+
Intercept	0.403	0.372	0.313	0.003
	(9.3)[a]	(5.1)[a]	(3.0)[a]	(0.0)
DX	0.042	0.017	0.083	−0.031
	(7.8)	(2.0)[b]	(6.2)[a]	(−2.0)[b]
DWXQ	−0.001	−0.0003	−0.002	0.000
	(−29.6)[a]	(−1.9)[b]	(−9.0)[a]	(0.5)
SEX	−0.136	−0.267	0.056	0.026
	(−1.9)[b]	(−2.7)[a]	(1.8)[c]	(0.9)
RES	0.123	0.198	0.153	0.124
	(9.5)[a]	(10.1)[a]	(6.7)[a]	(4.4)[a]
TCH	0.366	0.532	0.300	0.168
	(29.4)[a]	(26.9)[a]	(14.1)[a]	(7.4)[a]
MOBIL	−0.027	−0.048	0.019	0.022
	(−2.6)[a]	−(3.8)[a]	(0.9)	(0.7)
AGEPHD	−0.008	−0.007	−0.009	0.006
	(−7.4)[a]	(−3.2)[a]	(−2.9)[a]	(2.4)[a]
XPR	0.009	0.001	0.017	−0.003
	(7.9)[a]	(0.5)	(5.3)[a]	(−1.2)
CWA	0.055	0.001	0.056	0.192
	(6.4)[a]	(0.1)	(3.6)[a]	(9.9)[a]
SEX · *DX*	0.024	0.040	—	—
	(1.9)[b]	(2.1)[b]		
R^2	.14	.16	.09	.06
F	208.9[a]	134.0[a]	40.9[a]	16.8[a]
N	13,310	7152	3819	2214

[a]significant at the .01 level
[b]significant at the .05 level
[c]significant at the .10 level

motions than males, even though they did not receive significantly greater salary increases, as was shown in Table 1. Females received rank promotions on par with males within the older age cohort.

This evidence suggests that rank promotions are given to females as substitutes for salary increases within the middle-age cohort. It also suggests that Johnson and Stafford's expectation of an eventual narrowing of female-male salary differentials is more accurate for the narrowing of rank differentials. It supports Astin's conclusion (p. 10) that females claim more salary discrimination than discrimination in promotions. However, it does not support Astin and Bayer's (p. 199) or Blaine Mercer and Judson B. Pearson's (p. 266) conclusions that males and females have equal rates of promotion.

The higher promotion rate for middle-age females may be due to males having already achieved high ranks. In order to control for this and to determine whether there are any particular ranks out of which females have difficulty in being promoted, the three age cohorts were divided into subsamples designated by the individual's rank in 1960. Regressions were run on each of these subsamples and the results are shown in Table 3. Only the regressions whose coefficients of determination were significantly different from zero are shown. The correlation between *SEX* and *SEX* · *DX* was so great among associate and full professors that the interaction term was omitted from the regressions for these ranks. Promotions of full professors were promotions to deanships. Since a deanship is an administration position it would require, in many cases, a change of work activity. For this reason, *CWA* was omitted from the regressions on the subsample of full professors.

The effect of being female is to reduce sig-

TABLE 3—*OLS* REGRESSIONS EXPLAINING THE DICHOTOMOUS VARIABLE RANK PROMOTION BETWEEN 1960 AND 1966 BY AGE COHORT AND RANK IN 1960

	Instructor			Assistant			Associate			Full		
Variable	-40	40–50^a	50+	-40	40–50	50+	-40	40–50	50+	-40	40–50	50+
Intercept	0.822	—	1.847	0.948	—	—	0.920	1.046	—	−0.291	0.340	0.492
	(8.8)^b		(3.6)^b	(11.2)^b			(5.9)^b	(4.9)^b		(−1.0)	(2.2)^c	(3.4)^b
DX	0.020	—	0.012	−0.027			−0.084	0.039	—	0.013	−0.018	−0.124
	(1.8)^c		(0.1)	(−2.8)^b			(−4.6)^b	(1.3)		(0.4)	(−0.8)	(−5.5)^b
DXSQ	0.000	—	−0.001	0.001	—	—	0.002	−0.001	—	0.000	−0.000	0.001
	(−0.2)		(−1.3)	(5.7)^b			(5.8)^b	(−1.8)^c		(0.3)	(−.05)	(2.8)^b
SEX	−0.231	—	−0.836	−0.721	—	.	−0.154	−0.117	—	−0.233	−0.063	−0.034
	(−2.1)^c		(−1.9)^d	(−5.6)^b			(−2.3)^c	(−1.9)^d		(−1.5)	(−1.0)	(−0.7)
RES	0.112	—	−0.474	0.037	—	—	0.038	0.013	—	−0.086	0.006	0.004
	(3.0)^b		(−2.4)^c	(1.0)			(0.6)	(0.2)		(−1.1)	(0.2)	(0.1)
TCH	0.104	—	0.131	0.030	—	—	(0.053	0.039	.	−0.193	−0.111	−0.139
	(2.8)^b		(−0.9)	(0.9)			(0.9)	(0.5)		(2.8)^b	(3.1)^b	(−4.1)^b
MOBIL	0.017		0.277	−0.029		-	−0.010	0.015	--	0.120	0.232	0.184
	(−1.1)		(1.8)^d	(2.2)^c			(0.4)	(0.3)		(2.3)^c	(6.1)^b	(3.9)^b
AGEPHD	−0.002	—	−0.016	−0.000	—	--	−0.001	−0.014	--	0.020	0.003	0.009
	(−0.7)		(−1.5)	(−0.1)			(−0.3)	(−2.3)^c		(1.9)^c	(0.6)	(2.4)^c
XPR	0.001	-	0.034	−0.009	—	—	0.005	0.015	—	0.006	−0.001	0.008
	(0.4)		(3.0)^b	(3.1)^b			(1.0)	(2.2)^c		(0.5)	(−0.2)	(2.0)^c
CWA	0.019	—	−0.114	0.036	--	—	0.141	0.105	—	—	—	—
	(1.3)		(−1.1)	(3.0)^b			(6.4)^b	(3.1)^b				
SEX · DX	0.035	—	0.128	0.105	—	—	..	-	—	--	--	—
	(1.7)^d		(1.5)	(4.4)^b								
R^2	.07		.58	.04	.05	.03	.06	.02	.07	.05	.05	.06
F	3.2^b	--	3.7^b	10.1^b	1.5	1.0	10.6^b	2.3^b	1.5	3.1^b	8.7^b	10.0^b
N	423	43	38	2191	295	62	1474	883	184	463	1443	1244
Number female	34	8	6	77	46	22	37	51	45	7	42	88

[a] All individuals in this category received rank promotions
[b] significant at the .01 level
[c] significant at the .05 level
[d] significant at the .10 level

nificantly the chance of rank promotion for all the subsamples of age and rank below full professor for which regressions were significant. For young assistant professors and instructors, the coefficient for *SEX* · *DX* shows that an additional year of labor force participation during the 1960–66 interval resulted in a greater increase in the chance of rank promotion for females than males. However, full participation between 1960 and 1966 was not enough to offset the initial negative effect of being female. This is similar to the effect of additional labor force participation on salary increases of males and females in the young-age cohort. Apparently, females have to "prove" their professional interest before receiving rank and salary increases.

An iterative probit technique was also used to estimate the effects of the variables in Tables 2 and 3 on the probability of rank promotion. The *OLS* estimates of the effects of the explanatory variables on the probability of rank promotion were very close to the probit estimates for most variables. In particular, the estimates for the sex variables were quite close, although the probit estimates had consistently higher absolute values but the same signs.[7]

V. Summary

This study has used longitudinal data to compare life cycle salary increases and rank promotions for male and female academic scientists. The sample was divided into age groups roughly corresponding to the childbearing, childrearing, and childfree ages for the female. Percentage salary increases over the 1960–66 period were significantly greater for males than females only in the young-age cohort. The regressions suggested that an additional year of work experience over this period would result in a higher salary increase for females than males. Percentage

[7] For an explanation of probit analysis, see Henri Theil (pp. 628–32)

salary increases were equal for males and females in the middle- and old-age cohorts. Since other studies have shown that females have lower salaries than comparable males, the absolute salary increases for females would be lower than for males in these age cohorts, even though the ratio of males to female salaries remained constant over the period. A narrowing of the salary differential was not present for the older age cohorts, contrary to the prediction by Johnson and Stafford.

In the young-age cohort, females had a significantly lower chance of rank promotion than males. Within the middle-age cohort, females had a greater chance of rank promotion than males. In order to determine if this was due to females simply having farther to climb on the promotion ladder, the sample was subdivided by rank in 1960. Except for the promotion from full professor to dean, females had a significantly lower chance of promotion at all ranks and ages for which the *OLS* regressions were significant.

REFERENCES

H. S. Astin, "The Woman Doctorate in America: Demographic and Career Characteristics of Professional Women," unpublished rept., Commission on Human Resources and Advanced Education, New York 1966.

A. E. Bayer and H. S. Astin, "Sex Differences in Academic Rank and Salary Among Science Doctorates in Teaching," *J. Hum. Resources*, Spring 1968, *3*, 191–200.

David G. Brown, *The Mobile Professors*, Washington 1967.

N. M. Gordon, T. E. Morton, and I. C. Braden, "Faculty Salaries: Is there Discrimination by Sex, Race, and Discipline?," *Amer. Econ. Rev.*, June 1974, *64*, 419–27.

G. E. Johnson and F. P. Stafford, "The Earnings and Promotion of Women Faculty," *Amer. Econ. Rev.*, Dec. 1974, *64*, 888–903.

B. Mercer and J. B. Pearson, "Personal and Institutional Characteristics of Academic Scientists," *Soc. and Soc. Res.*, Apr. 1962, *46*, 259–70.

J. Mincer and S. Polachek, "Family Investments in Human Capital," *J. Polit. Econ.*, Mar./Apr. 1974, Part II, *82*, S76–S108.

Henri Theil, *Principles of Econometrics*, New York 1971.

U.S. National Science Foundation, *Doctorate Records File 1960–68*, (data tape).

The Earnings and Promotion of Women Faculty: Comment

By MYRA H. STROBER AND ALINE O. QUESTER*

In their recent article in this *Review*, George Johnson and Frank Stafford (J-S) attempt to downgrade the importance of discrimination in explaining the male-female wage differential among full-time faculty with a Ph.D. degree. They seek to substitute for the discrimination explanation the following alternative: "that the differential is primarily generated by the market's reaction to voluntary choices by females with regard to lifetime labor force participation and on-the-job training" (p. 889). Needless to say, a debate on the causes of the faculty wage differential by gender is nontrivial, since the policy implications of the alternative explanations are quite different. As J-S note, "If one accepts the conclusion that over half of the academic salary differential by sex can be explained by the market's reaction to voluntary choices of females regarding on-the-job training, then the implementation of antidiscrimination policies can be reconsidered" (p. 902). We argue here, however, that no relaxation of antidiscrimination policies is warranted. While the economic analysis is interesting and carefully done, we find the J-S case for a supply side explanation of the wage differential unconvincing.

Employing a human capital model, J-S propose that because new women Ph.D.s expect, at some time during their careers, to drop out of the labor force to care for children, their expected differences in labor supply over the life cycle operate so as to make them likely to select their first positions at institutions with high starting salaries. And, since, according to J-S, high-starting-salary institutions offer fewer opportunities to develop additional human capital, the seeds of future wage differentials are sown early by women themselves. Then, the story continues, women Ph.D.s dig their grave still deeper by dropping out of the labor force to raise their children, thus causing their already lower endowment of human capital to "depreciate." When they return, ultimately, to fill full-time faculty positions, their wages are lower than those of their male counterparts. After several years of post-childrearing labor force experience, women "reacquire skills" (p. 895) and the male-female wage differential ceases to rise or may even narrow, though, of course, the effects of earlier deficits in the acquisition of post-Ph.D. human capital are never fully overcome.

One would suppose that in order to test this life cycle human capital explanation for wage differentials, J-S would furnish empirical data on at least some of the following: (a) initial job preference patterns of women Ph.D.s; (b) starting salary differentials between prestigious (i.e., human capital-building) institutions and other institutions; and (c) drop-out and reentry patterns of women in academic positions. However, no information is presented on any of these issues. Indeed it is curious that the data with which J-S chose to work are absolutely unsuited to testing their life cycle human capital hypothesis, since the National Science Foundation Register provides no information at all on registrants' work histories. Instead of testing a hypothesis, J-S have fallen into the logical error of affirming the consequent. They have essentially said: if full-time women faculty behave in the manner which we postulate, the life cycle human capital model would generate a particular pattern of earnings differentials over the life cycle. Since we observe this particular pattern of earnings differentials in the cross-section estimation of our model, women faculty must be behaving as we postulated.

*Assistant professors of economics, Graduate School of Business, Stanford University, and Department of economics, State University of New York-Cortland, respectively

Our comment discusses six issues: (i) evidence on dropping out of the labor force and returning to a full-time academic appointment; (ii) the relevant time horizon for decision making and the tradeoff between salary and human capital-building opportunities; (iii) the worsening of the gender earnings differential during the 5–15-year postdegree period, and its later improvement; (iv) the dearth of women faculty at prestigious institutions and the dearth of women administrators; (v) the relevance of the noncitizen-citizen salary differential to an explanation of the gender differential; and (vi) the admissibility of the J-S method of determining how much of the gender differential is due to discrimination and how much is due to differences in the post-Ph.D. acquisition of human capital.

I. Evidence on Dropping Out of the Labor Force

The women in the J-S sample were Ph.D.s employed in full-time faculty positions at American institutions in 1970. As noted, there is no information on how many of the women in that sample had ever dropped out of the labor force. Many women Ph.D.s who drop out are unable to return to academic positions at all; others, by preference or circumstance, are employed full time but not in faculty positions. The J-S story might explain why the earnings of these women are quite low, but these women are not in the J-S sample. Since the assumption that a nontrivial proportion of current full-time faculty have interrupted their careers is central to the J-S argument, we present some evidence on the question.

In a study based on a sample of some 2,000 women who received their doctorates in 1957 and 1958, Helen Astin found that 7 or 8 years after receipt of the doctorate 79 percent of the women had never interrupted their careers. And for those who had, the median interruption was only 14 months. Further evidence of the labor force dropout of women Ph.D.s comes from preliminary data thus far collected in a survey being carried out by the Committee on the Status of Women in the Economics Profession (*CSWEP*) of the American Economic Association (*AEA*). (See Strober and Barbara Reagan for a description of the sample.) Of 212 women Ph.D.s employed full time as academic economists in fall 1974, only 9 percent ever had a gap in employment of 6 months or more since receipt of the Ph.D. degree, and even in the group of women over 35, those with an employment gap of 6 months or more represent only 12 percent of the total. It appears likely that women with interrupted careers may well represent only a small fraction of full-time female faculty.

II. The Relevant Time Horizon and the Tradeoff Between Salary and Human Capital Building Opportunities

If one expects to drop out of the labor force after only a few years of work experience and to *remain out thereafter*, the wisdom of maximizing short-run monetary gain is incontrovertible. However, women who expect to work for only a few years during their adult lives do not generally take out four or more years to obtain a Ph.D. Thus, the expectation for women Ph.D.s in that probably small group who do intend to drop out, is, ultimately, to return to work. Women Ph.D.s are therefore not likely to be interested simply in their initial salaries, but rather in their lifetime earnings. And, perhaps most important, those who expect to drop out are interested in assuring that they can reenter the academic world when they wish to do so. Thus, we argue, it is likely that women who expect to drop out, far from turning down an offer at a prestigious institution, would be particularly likely to accept it, as much for reasons of building contacts as for building skills. Any woman who expects one day to face a reentry situation knows that *whom* she knows will be at least as important as what she knows. Moreover, having once taught at Yale or M.I.T. can be as important a signalling device for a woman as for a man, even if the woman is signalling after a several year work hiatus.

J-S present no evidence, nor do we have any, on the number of women who turn down job offers at prestigious institutions. There are very few women in top departments, but whether this is a result of supply- or demand-side factors, or their interaction, remains unclear. However, it

does strain the imagination to picture women deliberately turning down offers at prestigious institutions because of low salaries. We have no data on starting salaries, but a brief investigation of American Association of University Professors' salary data for 1972–73 for institutions having the top-rated economics departments (see Alan Cartter) reveals that three of those seven institutions (M.I.T., Stanford, and the University of Chicago) had pay scales for assistant professors which were between the 80th and 95th percentiles. Among the seven schools, only Yale and Princeton paid particularly low salaries. Yet, there is no evidence that M.I.T., Stanford, or the University of Chicago had more women assistant professors than Yale or Princeton. Moreover, it is instructive to compare assistant professors' salaries among some of the schools in the Boston area. (See Table 1.) A woman who wished to teach at the assistant professor level in the Boston area and who had an offer from M.I.T. could, on the average, have topped that annual salary in 1972–73 only at Boston University (by $100). At Wellesley, Tufts, Brandeis, and Boston College, assistant professors had not only less prestige but also lower average salaries. In addition, they may have had fewer consulting and/or summer research salary opportunities. Clearly the existence, size, and effect of a tradeoff between starting salary and human capital-building opportunities or between initial job and consequent reentry possibilities need more research.

TABLE 1—SALARY FOR ASSISTANT PROFESSORS AT SELECTED INSTITUTIONS, 1972–73

Institutions Having the Top Seven Economics Departments[a]	Salary (In Thousands of Dollars)
Harvard	14.6
M.I.T.	15.2
Yale	13.6
Stanford	15.7
University of California, Berkeley[b]	14.5
Princeton	14.0
University of Chicago	14.7
Selected Institutions in the Boston area	
Boston College	14.9
Boston University	15.3
Brandeis	14.4
Tufts	13.7
Wellesley	14.9

Source: *AAUP Bulletin*
[a]As defined in Cartter.
[b]Data are for the entire University of California.

III. The Initial Worsening of the Gender Wage Differential and Its Eventual Improvement

For all six disciplines studied by J-S, the male-female wage differential becomes increasingly unequal as years of potential post-degree experience goes from 0 to 15 years. J-S maintain that this is so because these are the years when, for women, "child care is most prevalent" (p. 895). It is difficult to see how the prevalence of child care explains the increase in the gender wage differential among full-time faculty. Those women who are working part time or who are out of the labor force entirely during these years clearly cannot affect the full-time faculty gender differential, which J-S are studying. It may be that J-S mean to imply that the widening wage differential is explained by increasing productivity differences among full-time faculty in instances where women faculty have young children. However, it would appear that in keeping with their life cycle human capital hypothesis J-S would wish to point to the *reentry* of women during years 5–15 as being responsible for the increased wage differential.

Nonetheless there are alternative explanations for the worsening of the differential. Discrimination may simply worsen over time. Alternatively, the *effects* of discrimination may worsen over time as, for example, women invest less and less in themselves as they become increasingly discouraged by discrimination. J-S suggest that discrimination alone should not result in women investing less in themselves because "although the returns to investing in an additional unit of human capital are lower for a group which is discriminated against, the opportunity costs of investment are correspondingly lower" (p. 890).

If the nature of discrimination is simply a lower rental price per unit of human capital, then the above argument is correct. However, the nature of discrimination against women faculty

members does not appear to be exclusively of this variety. Rather, discrimination also consists of paying women faculty a price which bears only an irregular and unpredictable relationship to their human capital. For example, in four of the six academic fields which J-S studied, "females who received their doctorates from relatively prestigious schools do not have significantly higher salaries than females who attended other graduate schools" (p. 896) Indeed, in one of the fields (economics), the incremental return to attending a top-ranked graduate school was *negative*. J-S attribute these results to the fact that women faculty do not obtain the large amounts of postdegree experience required to yield a significant return on a prestigious degree. However, in their recent article, J-S find slightly higher earnings for economists with prestigious degrees even after only 5 years of postdegree experience (1974a, p. 558). Moreover, for the human capital explanation to fit, one would have to make the unrealistic assumption that women in two of the six fields somehow acquired more postdegree human capital than the other women And to explain the negative return would require some even more peculiar behavior postulate, although this negative return may be a statistical problem resulting from the age distribution as well as the small size of the female economist sample.[1]

Where the nature of discrimination is not simply a lower rental price per unit of human capital but a highly uncertain return, the opportunity costs of additional investments in human capital are *not* equal to the marginal returns. In a recent article David Levhari and Yoram Weiss inquire into the effects on human capital investment where earning power in the first period is random. They find that, given such an assumption, it is "in general impossible to determine whether the expected return on investing extra time in human capital . . . is above its expected alternative cost" (p. 960). We find no compelling evidence to dismiss discrimination as an important force in worsening the gender wage differential during the 5–15 year postdegree period.

One of J-S's interesting findings is that at "advanced years" of postdegree experience the male-female salary differential either grows at a much slower rate (anthropology, mathematics, biology) or narrows (economics, sociology, physics). Their explanation for this phenomenon is that by this time the women who have reentered the labor force have "reacquired" their skills We should like to suggest that discrimination in promotion practices could also yield J-S's result, since promotion often carries with it a substantial salary increment. If men tend to be promoted in the early years of their careers and women in the later years as the J-S data suggest (p. 896), then these differential promotion practices can explain not only the more rapid increase in women's earnings in the later years, with the consequent decline in the female-male wage differential, but also the growth of that differential in the early (fifth through fifteenth) years.

IV. The Dearth of Women Faculty at Prestigious Institutions and the Dearth of Women Administrators

J-S contend that if one wishes to employ a discrimination hypothesis to explain why women are less frequently on the faculties of prestigious institutions one needs to assume that the faculties of the Harvards and Yales of this world harbor stronger prejudices against women than do other institutions. Although such an assumption is plausible, it is not required for the

[1] Further evidence on the differences in returns to men and women faculty comes from a recent study by M. G. Darland et al. who looked at the effect on earnings of the number of articles and books written Articles and books may be viewed as a measure of productivity or as a measure of investment in human capital In either case, using data collected by the Carnegie Commission, the authors found that in the salary regressions for Research Universities Type I not only were the coefficients on the variables "number of articles" and "number of books" smaller for women than for men but also their significance varied: the number of articles written was significant in explaining male earnings in five of the six fields investigated, but significant for women in only three fields Similarly, number of books written was significant for men in four of the six fields, but significant for women in only two fields.

applicability of the discrimination hypothesis. As long as the best schools get first choice of the distribution of talent, the faculties of prestigious institutions could be the least discriminatory toward women and still have the smallest proportion of women faculty.

If the best schools choose first, or in effect choose first as candidates await decisions from the best schools before acting upon other offers, then only the best schools are able to select from the entire group of job market candidates. Assuming that among the candidates the distributions of ability by sex are identical, if the faculties at the best schools have even a *slight* preference for male faculty members, they will end up selecting a higher proportion of males than is contained in the distribution. This means that when the next-best schools begin their selection, the distribution of new Ph.D.s is different in two ways: 1) there are proportionally more women in the distribution than there were previously; 2) the distribution of ability by sex is now not identical; the females who are left are now slightly more able than the males. Thus, although the faculty in the less prestigious institutions may be more prejudiced against women, the *price* they must pay for their prejudice is much greater. If the taste for discrimination is uniform among schools, therefore, less prestigious schools are likely to end up with more women on the faculties. And, depending on the size of the differences in tastes and price among schools, this result may also hold even where the taste for discrimination is higher at nonprestigious institutions.[2]

[2]One might ask why discrimination at the more prestigious schools has taken the form of exclusion of women Why haven't prestigious institutions hired women and simply paid them less? It may well be that faculty members in high status institutions have a greater than average concern with the maintenance of prestige To the extent that high salaries and high prestige are viewed as reenforcing, the notion of employing faculty at rates lower than the established norm for a given rank might be particularly unpalatable to academics in high status institutions Moreover, a preference for wage differentials might be especially weak in a nonprofit setting where at the hiring (i e., departmental) level, faculty positions rather than salary dollars are often seen as the scarce resource.

The explanation which J-S offer for the dearth of women academic administrators is that administrative positions require the building of specific human capital and that "if higher turnover of women is expected, both women and employers will have incentive to invest in smaller amounts of specific human capital" (p. 897). We should first like to point out that even if women do tend to drop out of the labor force to raise children this does not mean that the turnover rate (separations rate) is higher for women than for men. Men also leave their jobs, albeit possibly for different reasons. Again, of course, the size of the male and female separations rates can be empirically examined. But even more importantly, it is unlikely that either males or females will be asked to do, or be interested in doing, much administrative work early in their careers. Faculty generally wait until after age 40 to begin investing in whatever "specific" human capital is required for administration. Women over 40 who are teaching full time are unlikely to be greater training risks than their male counterparts. It is our impression that women are absent from administrative posts not because they turn down opportunities to fill them or because they lack the relevant "know-how" about the operation of their departments or universities, but because they rarely receive the opportunity to "train" for these posts And the J-S evidence on the very substantial earnings increment received by those women who do administer only serves to strengthen our impression.

V. The Noncitizen-Citizen Wage Differential

In their attempt to show that discrimination is not a viable explanation for the male-female faculty wage differential, J-S test to see whether the wage differential between citizen and noncitizen faculty follows the same pattern over time as the male-female differential. The theory is that if male academics discriminate against women, then we might find that they discriminate against noncitizens in the same way. That is, if the two differentials do not follow the same

pattern, then discrimination is less likely to be responsible for the gender differential. The difficulty with this reasoning is that while noncitizens, if they are discriminated against, can become citizens, women cannot become men. Thus, elementary price theory would lead one, a priori, to expect different patterns over time in the two differentials.

In fact, J-S find that among biology professors at the top 20 graduate schools the noncitizen-citizen earnings differential improves over time, from .722 at 0 years of postdegree experience to unity after 10 years of such experience (p. 900). We suspect that the reason for the decrease in the differential is that the types of individuals in the noncitizen group at 0–5 years of potential postdegree experience are completely different from those in the group at higher potential experience levels, i.e., that the cross-section pattern of noncitizens' earnings bears little relationship to the life cycle pattern of noncitizens' earnings. We expect that at 0–5 years of postdegree experience J-S are observing foreigners who have remained in the United States after finishing their Ph.D.s. These individuals may face discrimination. However, if they plan to remain in the United States they would most likely become citizens. Otherwise they would return to their countries, and would not be included in the J-S sample. After 10 or more years of postdegree experience, the noncitizen group in the J-S sample are probably Canadians or distinguished visiting professors from universities outside the United States, neither of whom are likely to face discrimination.

VI. Admissibility of Assigning Weights to the Relative Roles of Discrimination and Human Capital Building

After reading J-S's explanation of wage differentials, the uninitiated might suppose that since women are supposed to rent out a higher fraction of their human capital when they first enter the job market, that the starting salaries of women Ph.D.s would be greater than the starting salaries of male Ph.D.s. However, this is not the case. According to Tables 3 and 11 in the J-S article, new women Ph.D.s earn 4–11 percent less, (7 percent less on the average) than their male counterparts. J-S choose to call this 7 percent the "discrimination coefficient." We should like to point out that if J-S are indeed correct, and that women do in fact rent out a higher fraction of their human capital on their first job, then 7 percent is merely the lower bound on the discrimination coefficient.[3]

But even if 7 percent were the correct discrimination coefficient for new Ph.D.s, there is no reason to assume, as J-S do, that this coefficient remains constant over time. (J-S assume that the entire worsening of the female-male wage differential over time is caused by gender differences in the post-Ph.D. acquisition of human capital.) Discrimination may increase over time if departments are more even-handed with respect to hiring than they are with respect to the speed of promotions and concomitant salary increases. There may be greater age discrimination against women than against men so that older women may suffer the effects of both sex and age discrimination. As noted earlier, the *effects* of discrimination may be cumulative as discouragement becomes progressively greater even if those doing the discriminating do not worsen their practices. Finally, new female Ph.D.s may face less discrimination today than did women in the past. If this is the case, inferring life-cycle earnings patterns from the cross-sectional earnings patterns is especially misleading.

[3]Taking account of supposed sex differences in the amount of human capital rented out, the discrimination coefficient may be calculated as follows: Let X = amount of human capital; Y = average annual male salary; α = fraction of human capital rented out by men; β = fraction of average annual male salary earned by women. If we assume that women rent out all of their human capital, then women's salary per unit of human capital is $\beta Y/X$; men's salary per unit of human capital rented out is $Y/\alpha X$. If we take the ratio of these terms we obtain $\alpha\beta$. The discrimination coefficient then is $1 - \alpha\beta$. For example, if men rent out 90 percent of their capital and women 100 percent, while women earn 93 percent of what men earn, then the discrimination coefficient is not 7 percent, but 16 percent.

REFERENCES

Helen S. Astin, *The Woman Doctorate In America,* New York 1969.

A. E. Bayer and H. S. Astin, "Sex Differentials in the Academic Reward System," *Science,* May 23, 1975, *188,* 796–802.

Alan M. Cartter, *An Assessment of Quality in Graduate Education,* Washington 1966.

M. G. Darland et al., "Application of Multivariate Regression to Studies of Salary Differences Between Men and Women Faculty," unpublished paper, Univ. California, Berkeley 1974.

G. E. Johnson and F. P. Stafford, (1974a) "Lifetime Earnings in a Professional Labor Market: Academic Economists," *J. Polit. Econ.,* May/June 1974, *82,* 549–70.

——— **and** ———, (1974b) "The Earnings and Promotion of Women Faculty," *Amer. Econ. Rev.,* Dec. 1974, *64,* 888–903.

D. Levhari and Y. Weiss, "The Effect of Risk on the Investment in Human Capital," *Amer. Econ. Rev.,* Dec. 1974, *64,* 950–63.

B. R. Reagan, "Two Supply Curves for Economists? Implications of Mobility and Career Attachment of Women," *Amer. Econ. Rev. Proc.,* May 1975, *65,* 100–07.

M. H. Strober, "Women Economists: Career Aspirations, Education and Training," *Amer. Econ. Rev. Proc.,* May 1975, *65,* 92–99.

American Association of University Professors, *Bulletin,* Summer 1973, *59,* 215–47.

The Earnings and Promotion of Women Faculty: Reply

By GEORGE E. JOHNSON AND FRANK P. STAFFORD*

The comments by Stephen Farber and Myra Strober and Aline Quester (S-Q) on our paper are very different in purpose and nature. The former is a utilization of longitudinal data to test some of our findings based on a single cross section; the latter is an expression of skepticism concerning our basic conclusion. This conclusion was that it is likely that slightly over half of the academic salary differential by sex (in 1970) would have existed even in the absence of direct labor market discrimination because of historical differences between the sexes in patterns of work attachment.

Before dealing with the substance of the comments, let us make two general points. First, in our paper (1974b) we did not maintain that direct labor market discrimination in the academic market place was unimportant; such a conclusion would be unwarranted on the basis of the analysis. Second, it should be noted that as a result of government pressures on the academic sector during the past five years our estimates for 1970 are likely to overestimate the currently prevailing total male-female salary differential and of the proportion of that differential attributable to discrimination.

In responding to the comments it is simpler to organize the discussion in terms of the six issues raised by S-Q and then turn to Farber's comment.

I. Labor Force Participation

The central point of our paper is that the household division of labor operates to increase home production of married women and to reduce their market activity. The resulting on-the-job training differences between men and women faculty, while smaller than the corresponding differences in the general labor market, do result in a salary differential between men and women faculty. Overall we find a 10–15 percent 9-month salary (wage) differential whereas a corresponding figure for all working women would be on the order of 50 percent (see Solomon Polacheck).

Although our *NSF* data do not contain information on marital status and family obligations, the data used by Helen Astin do have such information. There are two distinctive features of women doctorates which mitigate against obsolescence and limitation of professional development arising from reduced labor market participation. First, women doctorates are less likely to be married (55 percent versus 86 percent of a comparable age group) and second, "among women of comparable ages in the general population only 45 percent worked as compared with 91 percent of the women in this sample" (Astin, p. 58).

On the other hand, women doctorates were influenced by family in a manner parallel to that observed for women in other studies of the labor market. While full-time withdrawal from the labor force was rather uncommon, reduced participation in the form of part-time employment was more common and was most strongly influenced by marital status, husband's income, and presence of preschool children. Further, Astin demonstrated that, at least historically (1965 data), "the typical women doctorate in the sample spent about half her working time in teaching and about one-fourth in research. . . . [The] data showed that men doctorates, on the other hand, devoted a comparatively large proportion of their time to research and administration, (research, 41 percent; teaching, 31 percent; administration, 20 percent; other, 8 percent)" (pp. 63, 73).

*University of Michigan

Census data confirm the general pattern of part-time employment and intermittent participation. In 1960, 34.5 percent of women college professors worked 50–52 weeks per year compared with 58.3 percent of men, and during the sample week the average workweek for women was reported as 35.0 hours versus 42.3 hours for men. Intermittent participation can be inferred from data on the 1960 census "labor reserve" which for women college professors totalled 13,441 as compared with 38,367 in the labor force. This means that about 25 percent of women who had experience within the preceding ten years as college professors were out of the labor force at the time of the sample. The corresponding figure for men was 6 percent.[1]

New cohorts of women Ph.D.s are likely to have labor force participation patterns closer to those of men. The human capital interpretation of the historical record would thus imply a reduction of the earnings and promotion differentials between these new cohorts of men and women faculty. Further, as we noted in our paper, the on-the-job training interpretation does not rule out discrimination as an important factor but does imply that a substantial difference could arise even in the absence of overt or subtle forms of discrimination against women

II. Boston Area Salaries

The discussion of starting salaries in the Boston area provides limited evidence since there is no control for skill levels of new Ph.D.s in the two groups of schools. Perhaps the salary foregone to teach at schools with greater training options could be ascertained by examining differences in salary offers (and accepted salary offers) for the same individual This would effect a control for ability and graduate school training. In our (1974a) paper we noted that, in general, starting salaries of graduates of the most highly ranked Ph.D. programs are *below* those of persons from other programs. Since the salaries of those from the top ranked schools are higher later on, we infer that an implicit market for training options does exist and that women with family commitments will, on average, choose to "buy" less training.

III. The Widening of the Gender Differential

The presumption in our analysis is that although our data are restricted to those employed full time at the date of the survey, some of these women have previously experienced periods of part-time employment or labor force withdrawal. The probability of having experienced such a reduction in labor force activity and the cumulative magnitude of such reductions rises through years 0–15 of potential postdegree experience and is consistent with a human capital interpretation of the widening differential over this interval. We concede to S-Q that this earning pattern is also consistent with discrimination. In human capital models which include nonmarket activity in the objective function, reliance on the argument of "reduced opportunity costs for investment" arising from discrimination is weakened. This is because if women are not discriminated against in the nonmarket sector there would be an added incentive to engage in nonmarket activity and to build up the associated skills. In this setting market discrimination could result in a cumulatively larger earning differential by gender.

Though limited, the evidence we have on the promotion of never married women appears (at least to us!) to be consistent with the human capital interpretation. On the assumption that discrimination occurs independent of marital status, we find it important to note that never married women appear to have done far better than married women in terms of promotion to full professor.

IV. School Type and On-the-Job Training

Concerning the virtual absence of women

[1]See the U.S. Census, *1970:* "Labor reserve" is defined as comprising those qualified for employment by past experience but not currently seeking work

at highly ranked schools we agree (as in their fn. 2) that discrimination could occur more by quantity adjustments (i.e., fewer women) rather than through wages (i. e., lower salaries for women at a given school). The point in our paper was simply that type of school serves as a proxy for level of on-the-job training and given the historically observed labor force participation of academic women (as outlined by Astin), we would expect fewer women at the prestigious schools on this basis alone.

V. The Citizen-Noncitizen Differential

S-Q contend that our evidence on the erosion of the citizen-noncitizen differential is inadmissible because 1) noncitizens can become citizens, or 2) because noncitizens with advanced years of experience who face discrimination will leave, or 3) the presence of distinguished (high paid and with extensive postdegree experience) foreign visitors accounts for much of the apparent narrowing at advanced levels of potential postdegree experience. We note that if the noncitizens facing the greatest discrimination (those from places other than Canada or Europe?) have lower potential earnings in their home country, then they may still continue to work in the United States despite discrimination. Also since the *NSF* data base is restricted to members of *U.S.* professional societies, it is less likely to include distinguished *visiting* professors.

VI. Discrimination or Human Capital Differences?

In our (1974b) paper we used the initial starting salary differential as an estimate of the pure discrimination effect under the assumption that upon completion of the doctorate women and men have equal stocks of human capital and choose to rent them out in equal proportion. As we noted these assumptions are likely to be erroneous in two respects but the errors would work in opposite directions. Specifically, women would on average have accumulated less capital in graduate school if they expect greater family obligations, but such anticipated obligations should induce them to "rent out" a higher fraction of their human capital (or, alternatively, purchase fewer training options) when they first enter the labor force.

Whether the pattern that we observed will obtain in the future (even in the absence of antidiscrimination policies) will depend heavily on intrafamily choices of market and home production. These choices are much less amenable to direct policy action though of course they can be altered by, for example, changes in tax treatment of child care expenditures. Yet, under the division of labor-human capital approach to explaining lifetime earnings of women, we would anticipate smaller family sizes and a greater accommodation by husbands to the career goals of those married women faculty to result in reduced lifetime earnings differentials between men and women academics.

VII. Farber's Results

Taken at face value, Farber's results concerning changes in salaries from 1960 to 1966 offer *unusually* strong support for our estimate of the fraction of the salary differential caused by discrimination over the life cycle (see our Table 11). To see this, first assume that any salary difference in the initial job after receipt of the doctorate (say, age 30) is entirely due to labor market discrimination. Up to age 40, where postdegree experience increases from 0 to 10, the change in the logarithm of the ratio of female to male salaries is estimated by Farber to be $-.448 + .075\ DX$, where DX is the change in actual labor market experience over a 6-year period (see his Table 1). For women with a pattern of continuous work attachment $DX=6$, and the predicted effect on $log\ (S_f/S_m)$ of being female over a 6-year period is .002, or an increase of .03 percent per year. Between ages 40 and 50 the predicted change in $log\ (S_f/S_m)$ is 021 per 6 years, or an increase of .35 percent per year, and for ages above 50 the predicted change is $-.008$, or a decrease of 13 percent per year

These estimates actually imply that by age 50 (approximately 20 years of postdegree experi-

ence) the salary of a woman with continuous full time work attachment will have increased by 3.8 percent relative to that of a male professor. However, although we do not have the variance-covariance matrix of Farber's regressions, it is unlikely that this predicted change would be significant in a statistical sense. In any event, the results do appear to refute the hypothesis that discrimination against academic women increases with age.

Although Farber's results appear to support our basic argument of the effect of differential labor force participation, we have reservations about his analysis. First, the variable DX is subject to some degree of measurement error because it is restricted to certain values (i.e., 2, 4, 6) somewhat arbitrarily, does not distinguish work content, and does not distinguish people working "full time" for, say, 30 hours per week versus those working 60 hours per week. Thus, the estimated coefficients on the experience variables may be biased. Second, the form of Farber's regression is not entirely consistent with the model we employed. Specifically, the longitudinal equivalent of our equation (1) would have been (in his notation) $y = \beta_2 DX + \beta_3 DXSQ + \beta_4 SEX \cdot DX + \beta_5 SEX \cdot DXSQ + \beta_6 XPR \cdot DX \ldots$, instead of his form $y = \alpha_2 DX + \alpha_3 DXSQ + \alpha_4 SEX + \alpha_5 SEX \cdot DX + \alpha_6 XPR + \ldots$. Without actually estimating the version consistent with our model, it is impossible to judge how critical is this difference. However, the difference in specification may explain why Farber obtains a consistently positive estimate of α_6.

REFERENCES

Helen Astin, *The Women Doctorate in America,* New York 1969.

G. E. Johnson and F. P. Stafford, (1974a) "Lifetime Earnings in a Professional Labor Market," *J. Polit. Econ.,* May/June 1974, *82,* 549–70.

——— **and** ———, (1974b) "The Earnings and Promotion of Women Faculty," *Amer. Econ. Rev.,* Dec. 1974, *64,* 888–903.

S. W. Polachek, "Discontinuous Labor Force Participation and Its Effect on Women's Market Earnings," in Cynthia Lloyd, ed., *Sex, Discrimination and the Division of Labor,* New York 1975.

U.S. Bureau of the Census, *Census of the Population in 1970: Occupational Characteristics,* Subject Report PC (2)–7A, Washington 1971.

On the Length of Spells of Unemployment in Sweden: Comment

By Roger Axelsson, Bertil Holmlund, and Karl-Gustaf Löfgren*

In a recent article in this *Review* Nancy Barrett draws some strong conclusions concerning the length of spells of unemployment in Sweden and the United States. Her main point is that the average duration of a spell of unemployment is much higher in Sweden than in the United States, i.e., Swedish workers take longer to find jobs once they become unemployed[1] although they become unemployed less often than Americans. From this she concludes that the *U.S.* labor market is more efficient in the sense that job seekers are matched with vacancies at a higher speed.

The purpose of this note is to show the incorrectness of her assumptions and present more reliable estimates of the average duration of a spell of unemployment in Sweden. Our results indicate that the difference between the United States and Sweden in this respect is negligible.

I

How long does a person remain unemployed on the average? This simple question cannot be easily answered, despite the wealth of data available on the unemployed. The labor market statistics normally report the average duration of unemployment for those who are unemployed in a particular week. This is the cross-section measure reported by the Swedish *AKU* statistics, and this figure—which is a "second best measure" —is in fact used by Barrett. The "first best measure" which should be used is the average duration of a completed spell of unemployment.[2] Barrett uses some "unpublished material" to show that the second best measure is a good approximation of the first best measure. It can be proved that this is the case if the probability of remaining unemployed is constant regardless of the duration of the previous unemployment. However, if the probability of remaining unemployed increases with the duration of the previous unemployment—which seems to be the plausible assumption—then the second best measure overestimates the true length of a spell of unemployment.[3]

II

Barrett applies the following formula[4] to calculate the weekly flow of individuals into new unemployment (I_u) as a percent of the labor force

$$UN = I_u \cdot D_u \tag{1}$$

where D_u is the average duration of a spell of unemployment and UN is the rate of unemployment. In Barrett's calculations D_u is approximated by the available cross-section measure and UN is the official rate of unemployment in the Swedish *AKU* statistics.

We intend to show that Barrett's method of calculating I_u[5] implies that the total number of spells per year at three occasions are below the number of individuals who actually had experienced unemployment during the year, which evidently is absurd.

In our Table 1 the numbers in columns (1)–(3) are obtained from the Swedish *AKU* statistics. The data represent the total number of unemployed, i.e., include both sexes. By using equa-

*Department of economics, University of Umeå, Sweden

[1] As a matter of fact, Barrett's view is shared by several Swedish economists.

[2] In a demographic context the first best measure can be interpreted as the expected lifetime and the second best measure as the average age of the population.

[3] In other words it is possible to prove that if the death rate decreases with increasing age, then the expected life time is below the average age. For a proof see Tönu Puu, and Lofgren.

[4] The formula holds in a steady state.

[5] Our calculations are based on the correct data. In Table 1, p. 216, Barrett has some rather peculiar data for D_u. As far as we can see, only 5 out of 14 numbers are correct. This, however, does not affect her main conclusions.

TABLE 1—THE WEEKLY FLOW INTO NEW UNEMPLOYMENT (I_u), THE AVERAGE DURATION OF UNEMPLOYMENT FOR THOSE UNEMPLOYED (T_u), AND THE TOTAL NUMBER OF SPELLS PER YEAR

	T_u (1)	UN (2)	L (3)	I_u (4)	$N \cdot S$ (5)	N (6)	S (7)
1966	8.4	1.6	3792000	0.191	376621	310413	1.21
1967	10.3	2.1	3774500	0.204	400399	342966	1.17
1968	10 8	2.2	3822200	0 204	405459	343654	1.18
1969	12.4	1.9	3854900	0.153	306696	300542	1.02
1970	12.0	1.5	3913000	0 125	254345	291858	0 87
1971	13.5	2.5	3960800	0.185	381029	412776	0.92
1972	16.2	2 7	3969500	0.167	344711	396239	0.87

Source. Statistiska Centralbyrån (*SCB*) for columns (1)–(3), Arbetsmarknadsstyrelsen (*AMS*), 1973, p 33, for column (6)
L = the size of the labor force, yearly averages
N = the number of individuals experiencing unemployment during the year
S = the average number of spells per person
$N \cdot S$ = the total number of spells per year

tion (1), column (4) is obtained. The next step is to calculate the total number of spells per year. This is possible if we use the same formula as Barrett:

$$I_u = \frac{N \cdot S}{52 \cdot L} \tag{2}$$

The results are recorded in column (5). Column (6) represents the number of individuals experiencing unemployment during the year according to the annual retrospective surveys. We can now see from column (7) that if Barrett's estimate of the weekly flow into unemployment—built upon the second best measure (T_u)—is valid, it follows, for at least three years, that each unemployed has been unemployed less than once!

TABLE 2—THE NUMBER OF UNEMPLOYMENT PERIODS (THE RELATIVE DISTRIBUTION)

	Period $S = 1$	$S = 2$	$S \geqq 3$	Total
1966	74 7	14 8	10 6	100.0
1967	66 8	20 9	12 3	100 0
1968	62 5	21 3	16 3	100.0

Source Arbetsmarknadsstyrelsen (*AMS*) 1970, p. 10.

We shall presently show that her estimates of I_u for the remaining years 1966–68 must also be wrong.[6] Table 2 enables us to obtain a minimal estimate of the average number of spells per person. By assuming that the total number of individuals belonging to the class $S \geqq 3$ have experienced unemployment exactly three times per year we obtain what we below call S^{min}, i.e., a minimal estimate of the average number of spells per person. These minimal averages enable us, by using equation (2), to determine the minimal yearly inflows I_u, and finally equation (1) gives us maximal estimates of the first best measure D_u. These are shown in Table 3 for 1966–68. A comparison with Barrett's estimates shows that Barrett has $S < S^{min}$ and consequently her estimates of D_u exceed D_u^{max}.

TABLE 3—MAXIMAL ESTIMATES OF THE AVERAGE DURATION OF A SPELL OF UNEMPLOYMENT IN SWEDEN (1966–68)

	S^{min}	$N \cdot S^{min}$	I_u^{min}	D_u^{max}
1966	1.36	422162	0.211	7 58
1967	1.46	500730	0 250	8 40
1968	1 54	529227	0 265	8 30

Note: N is obtained from Table 1, col. (6)

III

We now intend to present some more plausible estimates of the average duration of a spell of unemployment in Sweden. The annual retrospective surveys do not give any estimate of the total number of unemployment spells per year,

[6]The estimate for 1969 is obviously wrong



FIGURE 1 THE WEEKLY FLOW INTO NEW UNEMPLOYMENT IN PERCENT OF THE LABOR FORCE

which Barrett seems to believe. The Swedish survey reports the number of individuals experiencing unemployment at least once during the year and, for some years, the distribution of unemployment spells according to Table 2. However, we do not know the distribution within the class $S \geqq 3$. As Robert Flanagan (1975, pp. 223–24) points out the same holds true for *U.S.* data on unemployment spells, which makes Barrett's calculations of the *U.S.* unemployment duration questionable.

The weekly flow into new unemployment can be calculated by taking the number of people unemployed one week or less in the cross-section distribution. The results of those computations are depicted in Figure 1.[7]

The average for both sexes is 0.3 percent which means that the weekly inflow into the stock of unemployed is about 12,000 persons (see Flanagan, 1973, p. 121). By using the cross-section unemployment duration—which is Barrett's method—we get much lower values of the size of the inflow (about 6–7,000 persons a week). This figure in per cent of the labor force is denoted I_u^* *and also depicted in Figure 1.*

[7]The Swedish *AKU* Survey has the class 0–2 weeks as its smallest break whereas the statistics on unemployment registered at the employment offices contain information on the number of people unemployed one week or less. Assuming that the annual relative distribution of registered unemployment within the class 0–2 weeks holds approximately also for the *AKU* Survey, we obtain an empirical estimate of I_u

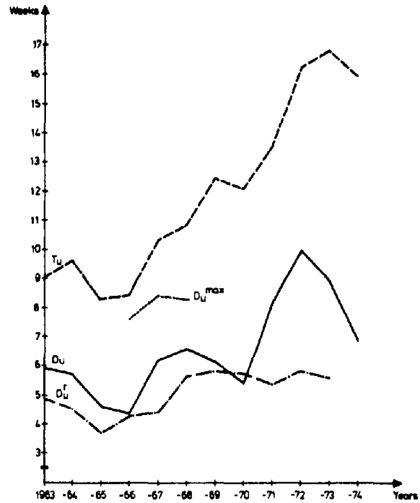


FIGURE 2 THE AVERAGE DURATION OF UNEMPLOYMENT 1963–74

By applying equation (1) it is now possible to get some rough estimates of the average duration of a completed spell of unemployment. The results are represented in Figure 2 (D_u). We also depict the cross-section average duration (T_u) in order to make comparison possible. There is obviously a considerable difference in magnitude; the average duration for completed unemployment spells is approximately only half as long as the cross-section figure as reported by the *AKU* Survey. Our estimates accordingly confirm the hypothesis that the probability of remaining unemployed increases with the duration of a spell of unemployment. As Hyman Kaitz (p. 12) has demonstrated this behavior holds for the United States, too.

There is a possibility that our figures to some extent underestimate the true length of an unemployment spell if the assumed distribution within the class 0–2 weeks is invalid. From Table 3 it is however quite clear that the cross-section values exaggerate the average duration of completed spells. The maximal estimates for the years 1966–68 are also depicted in Figure 2 (D_u^{max}).

It is possible to obtain an additional estimate of the average duration of an unemployment spell by taking the average annual weekly inflow of unemployed to the employment offices and the annual average of the stock of registered unemployment. By applying formula (1) we calculate D_u^r, which is depicted in Figure 2. We now get a still lower unemployment duration estimate. The fairly deep recession of the early 1970's has not affected the average duration of registered unemployment. The explanation is probably that the observed increasing propensity to register means that people who have been unemployed more than one week register at the employment offices and are then at first counted as unemployed less than one week. Such rising propensity to register means that the cross-section duration distribution of registered unemployment could be more skewed towards short-term unemployment.

IV

We have in this note attempted to show that the average duration of a completed spell of unemployment in Sweden is far below the cross-section estimates of unemployment duration. Our estimates indicate that the difference in this context between Sweden and the United States is pretty small, in fact Sweden appears to have a lower average duration of unemployment[8] as well as a lower rate of weekly inflow to the stock of unemployed. To conclude that these differences have no welfare implications seems to be exaggerated agnosticism. A large flow of new unemployment could to some extent reveal the importance of the "secondary labor market," for example. In fact, the problem of hard-core unemployment does not require a high rate of long duration unemployment. As has been pointed out by Robert Hall concerning the *U.S.* labor market, "some groups in the labor force have rates of unemployment that are far in excess of the rates that would accord with the hypothesis that the unemployed are making a normal transition from one job to another. Some groups exhibit what seems to be pathological instability in holding jobs. Changing from one low-paying, unpleasant job to another, often several times a year, is the typical pattern of some workers. The resulting unemployment can hardly be said to be the outcome of a normal process of career advancement" (p. 389).

[8]This conclusion is based on Barrett's figures for the United States, p. 216. It is somewhat strange that her estimates regard whites only and that she draws policy conclusions on such limited information.

REFERENCES

N. S. Barrett, "The U.S. Phillips Curve and International Unemployment Rate Differentials: Comment," *Amer. Econ. Rev.*, Mar. 1975, *65*, 213–21.

R. J. Flanagan, "The U.S Phillips Curve and International Unemployment Rate Differentials," *Amer. Econ. Rev.*, Mar. 1973, *63*, 114–31.

———, "The U.S. Phillips Curve and International Unemployment Rate Differentials: Reply," *Amer. Econ. Rev.*, Mar 1975, *65*, 222–25.

R. E. Hall, "Why is the Unemployment Rate So High at Full Employment?," *Brookings Papers*, Washington 1970, *3*, 369–402.

H. B. Kaitz, "Analyzing the Length of Spells of Unemployment," *Mon. Lab. Rev.*, Nov. 1970, *93*, 11–20.

T. Puu and K. G. Löfgren, "A Theorem on the Length of Spells of Unemployment," dept. econ., Univ. Umeå, mimeo. 1975.

Arbetsmarknadsstyrelsen *(AMS)*, "Arbetsmarknadens beteendemönster" (The Behaviour Pattern of the Labor Market), *25*, Stockholm 1973.

———, "Arbetslösheten under 1968" (Unemployment during 1968), *3*, Stockholm 1970.

Statistiska Centralbyrån *(SCB)*, "Arbetskraftsundersökningar" ("Swedish Labor Force Surveys," *AKU*), yearly averages 1966–72.

On the Length of Spells of Unemployment in Sweden: Reply

By Nancy Smith Barrett*

In their comment on an earlier exchange in this *Review* between Robert Flanagan and myself, Roger Axelsson, Bertil Holmlund, and Karl-Gustof Löfgren (A-H-L) dispute my findings that the average duration of unemployment in Sweden has, for the past decade, been longer than in the United States. Their argument is unconvincing, however, for several reasons. First, it is impossible to determine whether the bias in the Swedish retrospective survey data that they use to demonstrate their point stems from an overestimate of duration (as they assume) or of the number of persons experiencing unemployment. Second, while they "adjust" my estimate of duration of unemployment for Sweden, they fail to make a similar adjustment for the *U.S.* data. If there is an error in my estimating techniques, the bias is likely to be in the same direction for both countries since the unemployment surveys are similar and are likely to have similar biases. Finally, there is other evidence that the average duration of unemployment is longer in Sweden than in the United States.

A-H-L correctly observe that the duration concept used in the monthly labor market surveys in Sweden and the United States is the length of time an unemployed person has been unemployed, not the average length of a completed spell. Decomposing unemployment into turnover and duration as Flanagan and I did in our earlier studies requires estimating the average length of a completed spell, a statistic that is not available from the monthly surveys.

One source of data on the length of completed spells is the annual retrospective surveys conducted in each country that ask individuals to report their unemployment experience over the entire year. However, these surveys are subject to a number of biases associated with retrospective reporting and are generally considered less reliable than as a measure of unemployment than the monthly surveys.

A-H-L contend that my duration measures for Sweden would result in a situation in which the average number of unemployment spells per year at three occasions are below the number of individuals who experienced unemployment. However, their data on the number of persons experiencing unemployment are obtained from the Swedish retrospective survey that shows an even greater inconsistency in this respect than my estimates show. For the years 1966, 1967, 1968, and 1969 (the only years for which I have the full survey results available), data from the Swedish retrospective survey show an average number of spells of 0.886, 0.803, 0.745, and 0.955 respectively, using the formula:

$$S = \frac{U \times L \times 52}{N \times D}$$

where

S = average number of spells per unemployed person per year
N = number of persons experiencing unemployment in the year
L = average labor force (monthly survey concept)
D = average duration of unemployment (in weeks)
U = unemployment rate (monthly survey)

This means that in the Swedish retrospective survey either D, average duration of unemployment or N, the number of persons experiencing unemployment is seriously biased upward. A-H-L seem to assume implicitly that all the bias is in the duration component, but they show no evidence that the bias is not in N. I recognized this problem when I estimated duration for

*Professor of economics, American University, on leave as Deputy Assistant Director for Fiscal Analysis, Congressional Budget Office.

Sweden and used a lower duration estimate than that obtained from the retrospective survey. While my duration estimates for Sweden may still be high, a test of the consistency of my estimates should not use an estimate of N that is very likely to have a strong upward bias. I would suggest that a different (lower) estimate of N would show that in all cases the average number of spells per person would exceed one. Further, the revised duration estimates by the authors involve use of the same upwardly biased N which implies their duration estimates for Sweden are much too low (D being inversely related to N for a given unemployment rate).

Second, if it is true that an alternative method for estimating duration of unemployment is desirable, the same techniques should be applied to both the *U.S.* and Swedish data for making comparative statements. Since both the retrospective and monthly surveys are similar for both countries, any biases are likely to be in the same direction. For comparative purposes, therefore, it is not very useful to demonstrate that the estimate of duration for Sweden is too high unless it is also shown that the estimate for the United States is not too high.

Finally, other evidence can be brought to bear that indicates rather clearly that Swedes experience, on the average, longer spells of unemployment than American workers. First, if one compares the duration of unemployment reported for those unemployed in the monthly surveys, the average duration in Sweden for the period 1967–72 averaged 12.9 weeks compared with 9.5 weeks for the United States.[1] While both of these measures may overestimate the duration of completed spells (to the extent that the survival rate increases with the length of the unemployment spell) the relative difference between the measures might not be very much different if survival rate patterns in both countries are similar.

Further, Flanagan in our original exchange of articles, finds that there is no significant difference between the rate of long-duration unemployment (the percent of the labor force unemployed 15 weeks or longer) between the two countries. Since the *U.S.* overall unemployment rates are much higher than for Sweden, long-duration unemployment represents a smaller proportion of the total unemployment in the United States. This is borne out by Table 2 of my original paper that shows a much higher proportion of the unemployed experiencing spells of 13 weeks or more in Sweden than in the United States. Although this comparison uses the monthly survey concept of duration, the direction and magnitude of the difference is striking.

A final consideration that I found particularly striking in the Swedish data is the remarkable inverse relation between duration and turnover of unemployment in various cross sections—most notably across age groups and regions, and a lack of any distinct relationship in the American data. This led me to the conclusion that longer expected durations of unemployment in Sweden may discourage a certain amount of job mobility in response to market signals. This *captive worker effect* may result in a productivity loss in the Swedish labor market even though unemployment is relatively low. It is interesting to note that the A-H-L revised estimates of duration and turnover in Sweden show that between 1965–66 and 1971–72 the increase in unemployment that occurred was entirely due to increased duration, with the incidence of unemployment actually declining over the period.

Despite the ambiguities in their analysis, however, the Axelsson, Holmlund, Löfgren piece is important in its recognition of the importance of decomposing unemployment flows into turnover, that is, frequency of occurrence of unemployment, and duration, that is, average length per spell. These two components have different underlying determinants and hence different implications for remedial policies. Any international labor market analysis should go beyond gross comparisons of unemployment rates and examine these components where possible. However, measuring these components, particularly the length of completed spells is difficult,

[1] I was not able to obtain the later Swedish surveys in the time allotted to prepare this note.

not only given the limitations of available surveys, but also because of the likely biases involved in retrospective reporting. Continued pursuit of improved estimating procedures can only add to our understanding of this important aspect of unemployment and enable us to pursue more effective labor market policies.

REFERENCES

R. Axelsson, B. Holmlund, and K.-G. Löfgren, "On the Length of Spells of Unemployment in Sweden: Comment," *Amer. Econ. Rev.*, Mar. 1977, *67*, 218–21.

N. S. Barrett, "The U.S. Phillips Curve and International Unemployment Rate Differentials: Comment," *Amer. Econ. Rev.*, Mar. 1975, *65*, 213–21.

R. J. Flanagan, "The U.S. Phillips Curve and International Unemployment Rate Differentials: Reply," *Amer. Econ. Rev.*, Mar. 1975, *65*, 222–25.

Earnings, Productivity, and Changes in Employment Discrimination During the 1960's: Additional Evidence

By JAMES E. LONG*

A recent article in this *Review* by Joan Haworth, James Gwartney, and Charles Haworth (hereafter H-G-H) presented some significant findings on the source and structure of improvements in the relative economic status of nonwhite males during the 1960's. Specifically, H-G-H concluded that approximately one-half of the increase in the nonwhite/white earnings ratio (*NWER*) during the 1960's was simply "... attributable to the exiting of older nonwhite workers with low relative earnings combined with the entry of younger, better-prepared nonwhites who have high relative earnings" (p. 167). The balance of the gain in relative nonwhite earnings was the result of a decline in employment discrimination against nonwhites and improvements in the relative productivity of nonwhites.

If correct, these findings have several important implications for the prospect of black and white earnings equality. First, they suggest that the effects of *past* discriminatory practices in both employment and the acquisition of human capital continue to reduce the earnings power of older black males still in the labor force. These past practices are an important source of *current* differences in the average earnings of blacks and whites in aggregate. Hence, they constrain the success of policies to achieve racial earnings equality.

This brief note presents some additional evidence consistent with the H-G-H conclusions. We are mainly concerned with their data in Table 3 (p. 164) on changes in the *NWER* within age cohorts between 1959 and 1969 The H-G-H hypothetical "identical productivity" *NWER* measures the nonwhite-white earnings gap caused by factors other than measured productivity variables, such as racial differences in occupational structure which are unrelated to productivity differences. Since the relative occupational distributions indirectly revealed by these hypothetical *NWER* underlie some of the major H-G-H conclusions, a more direct examination of changes in the occupational distribution of blacks and whites during the 1960's may prove a useful check on their findings.

A group's index of occupational status can be calculated by weighting the proportion of the group employed in an occupation by the mean earnings for the occupation and summing across major occupational categories.[1] The higher (lower) a group's index, the greater the proportion of the group in higher (lower) paying occupations. Therefore, the ratio of nonwhite to white occupational status (*NWOS*) will measure solely racial differences in the distribution of workers among occupations. The higher (lower) the *NWOS*, the more (less) favorable the occupational structure of blacks relative to whites, *ceteris paribus*.

Estimates of the male *NWOS* for age cohorts in 1959 and 1969 are presented in Table 1, along with the corresponding *NWER*.[2] In aggregate, the *NWOS* increased by 10.2 percent during the

*Assistant professor of economics, Auburn University

[1]This index of occupational status equals $\Sigma_i p_i^j e_i^j$, where p_i^j is the proportion of the jth class (such as age) of the group in the ith occupation, and e_i^j is the mean earnings in the ith occupation of all persons in the jth class

[2]The *NWOS* was calculated directly from published census data for all cases except the age cohorts 35–44 and 45–54 in 1970. Because mean earnings by occupation in 1969 were available only for the combined age class 35–54 years, separate earnings data for these two cohorts had to be estimated using the fact that age 35–54 earnings (X) are a weighted average of age 35–44 earnings (Y) and age 45–54 earnings (Z), i e., $X = \alpha Y + \beta Z$, where α and β are the fractions of males age 35–54 with earnings in 1969 who are 35–44 and 45–54 years of age, respectively Assuming that the lifetime earnings profile did not change during the 1960's, that is Y/Z was equal to the 1959 value (k), we derived estimates of Y and Z where $Y = kX/k\alpha + \beta)$ and $Z = X/(k\alpha + \beta)$.

TABLE 1—ESTIMATED CHANGES IN THE *NWOS* AND *NWER* OF MALES IN AGGREGATE AND WITHIN COHORT GROUPINGS BETWEEN 1959 AND 1969

	NWOS			*NWER*		
	1959	1969	Percent Change	1959	1969	Percent Change
Aggregate						
Employed Males age 14 and over[a]	72 5	82.9	+10 2	70.8	79.0	+11.6
Cohort Group						
18–24 in 1970	[b]	93 4	[b]	[b]	[b]	[b]
18–24 in 1960 (25–34 in 1970)	82.3	87.9	+6 8	[b]	[b]	[b]
25–34 in 1960 (35–44 in 1970)	78 6	81.4	+3 6	73.7	77 9	+5 7
35–44 in 1960 (45–54 in 1970)	74 3	76 7	+3 2	70.8	74 0	+4.5
45–54 in 1960 (55–64 in 1970)	68 8	75.5	+9 7	65.1	70.6	+8 4
55–64 in 1960	67.4	[b]	[b]	[b]	[b]	[b]
Unweighted mean[c]	76.0	80 4	+5 8	69 9	74 1	+6.0

Source. *NWOS* calculated with 1960 and 1970 Census data from *Detailed Characteristics*, *Occupation by Earnings and Education*, and *Earnings by Occupation and Education*. *NWER* taken from H-G-H, Table 2, p 162, and Table 3, p 164

[a]For 1969 ratios, males age 16 and over

[b]Not applicable

[c]Excluding 18–24 in 1970 and 55–64 in 1960

decade. Gains in the *NWOS* within cohorts were generally much smaller; 5.8 percent on average. The *NWOS* was negatively related to age, indicating an improvement in the relative occupational status of nonwhites with each successively younger age cohort. Not surprisingly, the pattern of the *NWOS* over time closely parallels that exhibited by the *NWER*.

Changes in this hypothetical *NWER* reflect only changes in the earnings functions of blacks and whites, since productivity factors are held constant *in the calculations*. Our estimates of the *NWOS* in 1959 and 1969 suggest that shifts in the relative occupational structure of blacks were a major cause of these earnings function changes. By 1969 young blacks and whites were facing similar opportunities to enter various major occupations (the *NWOS* for males 18–24 in 1970 was 93.4). The *NWOS* for labor force entrants during the 1960's (persons 18–34 in 1970) was substantially greater than the *NWOS* for those who probably exited from the labor force during this period (persons 55–64 in 1960). The relative occupational position of middle-aged blacks showed only minor improvement over the 1960's. However, the *NWOS* of older workers (age 45–54 in 1960) increased substantially (9.7 percent), as did the corresponding *NWER* (8 4 percent) H-G-H contend that this large increase was achieved primarily through the least productive blacks with low earnings leaving the labor force between 1959 and 1969, and not through increased upward mobility among older black workers.

Changes during the 1960's in the actual number of older black males employed in various occupations are consistent with this hypothesis. The index of occupational status for older blacks increased during the 1960's because large declines in the absolute number of blacks in low-paying farm and laborer jobs reduced the proportion of blacks in such jobs (see Table 2). At the same time, an increased representation of older blacks in higher-paying jobs (like professional-technical, clerical, and craft) resulted because declines in black employment in these occupations were much smaller. This mechanism—the largest declines in black employment during the 1960's occurring in occupations yielding the lowest earnings—was the primary source of the

TABLE 2—CHANGES IN BLACK EMPLOYMENT DURING THE 1960'S FOR THE COHORT GROUP 45–54 IN 1960 (55–64 IN 1970)

	Number of blacks employed (000's)			Percent of blacks in the occupation		
	1960	1970	Percent Change	1960	1970	Percent Change
Professional and Technical	22	18	−18	2.67	3.34	+25
Farmers and Farm Managers	40	11	−72	4.85	2.01	−59
Managerial and Administrative	22	15	−32	2.67	2 83	+6
Clerical	31	27	−13	3.76	5 07	+35
Sales	9	6	−33	1 09	1 08	−1
Craftsmen	94	75	−20	11.40	14 20	+25
Operatives	190	133	−30	23 03	25 12	+9
Service Workers	124	113	−9	15.03	21 45	+43
Farm Laborers	50	26	−48	6.10	4.85	−21
Laborers (except farm)	172	102	−41	20.85	19 34	−7
All Occupations	825	529	−36			

Source: See Table 1

rise in the *NWOS* and *NWER* of older males between 1959 and 1969.[3]

Examining changes in the relative structure of nonwhite employment during the 1960's has provided some evidence consistent with the H-G-H conclusion that increased nonwhite/white earnings during this period were largely the result of declining labor force participation among older nonwhites. One major implication of this finding is apparent. Past discrimination in both employment and other areas retards our ability to promote earnings parity between blacks and whites. The chances of older black workers completely making up earnings losses due to past discrimination are slight, especially if this is to be accomplished without reverse discrimination against existing white workers. Thus, equal opportunity will lead to earnings equality only with the passage of time and the exit from the labor force of older blacks whose human capital and skill-building experience have been diminished by past discriminatory practices.

[3]To a lesser extent, the occupational status of older whites also increased during the 1960's because of their declining labor force participation Explaining these employment changes among older blacks and whites is not our concern in this note, but we might expect that changes in social security laws affecting the retirement age and poor health (see Herbert S. Parnes and Jack Meyer, and Harold S. Luft) would be causes of withdrawal from the labor force by these older workers.

REFERENCES

J. G. Haworth, J. Gwartney, and C. Haworth, "Earnings, Productivity, and Changes in Employment Discrimination During the 1960's," *Amer. Econ. Rev.*, Mar. 1975, *65*, 158–68.

H. S. Luft, "The Impact of Poor Health on Earnings," *Rev. Econ. Statist.*, Feb. 1975, *57*, 43–57.

H. S. Parnes and J. Meyer, "Withdrawal from the Labor Force by Middle Aged Men, 1966–67," mimeo, Center Human Resource Res., Ohio State Univ., Jan. 1971.

U.S. Bureau of the Census, *U.S. Census of Population: 1960, Detailed Characteristics*, PC(1)-D, U.S. Summary, Washington 1963.

———, *U.S. Census of Population: 1960*, Subject Reports, *Occupation by Earnings and Education*, Final Report PC(2)-7B, Washington 1963.

———, *U.S. Census of Population: 1970, Detailed Characteristics*, Final Report PC (1)-D1, U.S. Summary, Washington 1973.

———, *U.S. Census of Population: 1970*, Subject Reports, *Earnings by Occupation and Education*, Final Report PC(2)-8B, Washington 1973.

Excess Demand, Search, and Price Dynamics

By STEPHEN MCCAFFERTY*

In this paper, I analyze the behavior of demanders and suppliers and the allocation of goods in a single market in which the prevailing price initially generates excess demand. The basic model assumes that in the face of the existing shortage, demanders search over space for the available suppliers. At any given time, some fraction of potential sellers are active suppliers, i.e., have a quantity of the good to sell, and the remaining potential suppliers are sold out. Demanders know the average availability of the good, but at any given time they do not know the location of these supplies. They search for these supplies until they find an active supplier or they drop out of the market. The analysis shows how for a fixed price the interaction between this shopping activity and the quantity supplied would determine a steady-state solution for the amount of shopping and the fraction of sellers who are active suppliers.

An important objective of the determination of the optimal amount of shopping is to provide an analytical basis for the distinction between the quantity demanded and the quantity ordered, in the context of excess demand. The quantity demanded is the quantity which, if bought, would be consistent with the maximization of buyer objectives, given the price and any other constraints on buyer behavior. In contrast, Herschel Grossman's discussion defines the quantity ordered as a measure of buyers' active attempts to purchase the good. In the present analysis the quantity ordered specifically equals the total number of search attempts people make to purchase the good. If the market were clearing, and all the suppliers were open, then the good could be readily purchased on the first search attempt. In this case the quantity ordered would equal the quantity demanded. However, in situations of excess demand, when some stores are closed, prospective buyers must search for open stores. This consideration can cause the quantity ordered to be either greater or less than the quantity demanded. As one possibility, some individuals might make repeated search attempts to try to purchase the quantity demanded. Bent Hansen's important study of inflation emphasized this possibility of overordering. Alternatively, because the time spent searching is costly, and search may be fruitless, there is a disincentive to search. Consequently, some individuals might not search at all. If there are enough of these individually discouraged demanders, on average, people may make less than one search attempt. In this case the quantity ordered would be less than the quantity demanded.

The analysis also investigates the implications of buyers' shopping behavior on price dynamics. Since sellers observe only the quantity ordered, and not the quantity demanded, it is reasonable to assume that if sellers set prices, prices will respond to changes in the quantity ordered rather than the quantity demanded. The formal analysis therefore postulates that the rate of price change is proportional to the difference between the quantity ordered and the quantity supplied. This assumption is in contrast to the standard assumption that the rate of price change is proportional to the level of excess demand.

The main result of this analysis is that it allows for the possibility of the price overshooting the market clearing price. That is, in adjusting to a situation of excess demand, it is possible for the price to rise above the market-clearing price before again dropping back down to the market-clearing price. This possibility arises because even when supply temporarily exceeds demand not all stores have yet been stocked and so the quantity purchased may still be less than the quantity demanded. In this situation, the quantity ordered can exceed the quantity demanded by such an extent that it may also exceed

*Assistant professor of economics, Ohio State University. This paper is taken from my doctoral dissertation. The research was supported in part by a Research Fellowship at The Brookings Institution. I wish to thank my advisor Herschel Grossman for many helpful comments and suggestions. Valuable comments were also received from an anonymous referee. I, of course, take full responsibility for any remaining errors.

the quantity supplied. Therefore the price can continue to rise temporarily above the market-clearing price.

I. Shortages and Optimal Shopping

Assume that at any given time the representative demander knows the fraction of potential sellers who are active sellers—that is, he knows the fraction of shops which are open for business and the fraction which are sold out and closed. However, he does not know which shops are the open shops and which are sold out. In order to locate an open shop, he must engage in the time consuming process of search. A demander will search either until he locates an active supplier or until the marginal utility cost of continuing to search equals the expected marginal utility benefit of continuing to search.

Assume that the marginal utility cost of visiting one more shop is given by $a \cdot s$, where a is a positive constant and s measures both the amount of time already spent shopping and the number of shops already visited, on the assumption that the time cost of visiting one more shop is constant.[1] This assumption says that the marginal utility cost of shopping increases with the amount of time spent shopping. The rationale for this relationship is the decreasing marginal utility of leisure.

Next, assume that the expected marginal utility benefit of visiting one more shop is given by $\beta \cdot X$, where β is the known fraction of potential sellers that are active sellers and X measures the addition to utility which would result from purchasing the desired quantity of the good.[2] Shopping will continue either until the demander locates an active supplier or as long as

$$as < \beta \cdot X \tag{1}$$

Therefore, the maximum amount of time the individual is willing to spend searching is the maximum integer $\bar{s}$ for which

$$\bar{s} < \beta X / a \tag{2}$$

[1]An alternative and slightly more complex formulation would assume that the marginal time cost of shopping increases with the number of shops visited.

[2]Assume that any active supplier will sell the desired quantity; that is, we neglect the possibility that orders may only be partially filled. Implicit in the analysis is the assumption that the individual's level of utility is given by $Z - 1/2\, a s^2$. The value of Z is zero if he fails to make a purchase and X if he succeeds.

To simplify the mathematics the present analysis allows $\bar{s}$ to take on noninteger values. That is, the analysis assumes that the representative demander chooses

$$\bar{s} = \beta X / a \tag{3}$$

This value of $\bar{s}$ can never differ from the optimal integer value by more than one. If, after an amount of shopping $\bar{s}$, the demander has not located an active supplier, he abandons his search and becomes what may be called a "discouraged demander."

It will be useful first to compute the probability density function of search time under the assumption that the prospective buyer is willing to search forever. Since the probability that any search attempt will be successful is equal to β, the amount of time needed for successful search will be binomially distributed as

$$P(s = k) = \beta(1 - \beta)^{k-1} \tag{4}$$

Taking into account the optimal stopping rule given by equation (3), the probability that search will be successful q, is given by

$$q = \sum_{k=1}^{\bar{s}} \beta(1 - \beta)^{k-1} = 1 - (1 - \beta)^{\bar{s}} \tag{5}$$

or

$$q = 1 - (1 - \beta)^{\beta X / a} \tag{6}$$

The expected value of search time taking into account the possibilities of successful search and termination at $\bar{s}$ is given by

$$E(s) = \sum_{k=1}^{\bar{s}} k\beta(1 - \beta)^{k-1} + \bar{s}(1 - q) \tag{7}$$

This expression can be simplified to[3]

[3]This is easily proven. Multiplying equation (7) by $(1 - \beta)$ yields

$$(1 - \beta)E(s) = \sum_{k=1}^{\bar{s}} k\beta(1 - \beta)^{k} + \bar{s}(1 - \beta)^{\bar{s}+1}$$

Now note that equation (7) can also be written as

$$E(s) = \sum_{k=0}^{\bar{s}-1} (k + 1)\beta(1 - \beta)^{k} + \bar{s}(1 - \beta)^{\bar{s}}$$

Subtracting the first equation from the second permits derivation of

$$\beta E(s) = \sum_{k=1}^{\bar{s}} \beta(1 - \beta)^{k-1}$$

and equation (8) now follows directly from equation (5).

(8) $$E(s) = \frac{1-(1-\beta)^{\bar{s}}}{\beta} = \frac{q}{\beta}$$

Application of L'Hôpital's rule shows that

(9) $$\lim_{\beta\to 0} q = \lim_{\beta\to 0} E(s) = 0$$

(10) $$\lim_{\beta\to 1} \ = \lim_{\beta\to 1} E(s) = 1$$

Consider a market in which the representative demander would like to purchase an amount $d(\bar{P})$, $d' < 0$, where $\bar{P}$ is a fixed price which generates excess demand. A flow of N demanders enters the market per unit of time and each one implements the optimal stopping rule, given by equation (3).

The quantity demanded per unit of time will be given by

(11) $$D(\bar{P}) = Nd(\bar{P})$$

This is the quantity that individuals would purchase if they did not have to search; that is, if β were equal to one. The probability that a demander will make a successful purchase given by equation (6), times the quantity demanded,

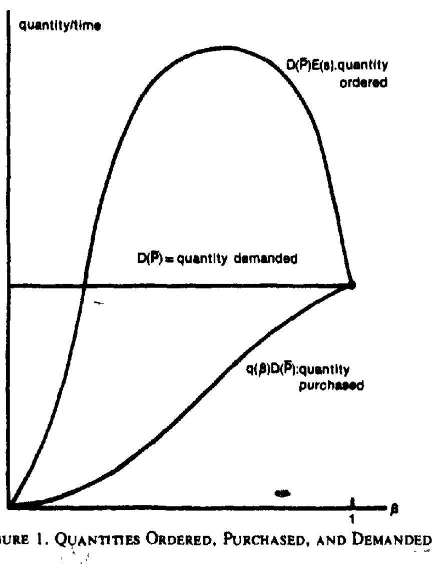


FIGURE 1. QUANTITIES ORDERED, PURCHASED, AND DEMANDED

indicates the expected quantity purchased. The expected value of search time given by equation (8) indicates the average number of attempts a demander makes to purchase the commodity. This number times the quantity demanded may be identified as the quantity ordered.[4] The aggregate quantities ordered, purchased, and demanded are depicted in Figure 1 as functions of β. Note that the quantity ordered may be either greater or less than the quantity demanded. The quantity ordered will be greater than the quantity demanded when many demanders are searching at more than one shop. The quantity ordered can be less than the quantity demanded when a significant number of demanders do not bother to enter the market at all due to the low expected returns to search.

II. The Inventory Process

A variety of patterns of distribution of the commodity amongst the potential suppliers can be consistent with the above assumptions regarding demander behavior. One possibility is that the potential suppliers are sales distribution centers for an exogenous producer. This producer acts in such a way as to keep some of his outlets stocked at all times by synchronizing deliveries to them with their sales. The number of outlets the producer can keep stocked will be determined by rate of sales that the outlets experience such that the number of outlets stocked times the rate of sales equals the flow of output produced. This distribution will be such that either the outlets supplied are rotated or the individual buyers do not re-enter the market often enough to get to know when specific outlets are stocked.

Another possibility would be production by

[4]The probability that a demander will make a successful purchase q, and the expected value of search time $E(s)$ are assumed to be independent of the price $\bar{P}$. Since the utility value of the purchase X will tend to be higher at a lower price, this assumption is not really accurate (see equations (6) and (9)). However, this consideration is of no consequence in the present context of a fixed price, and also does not substantially change the results of Section IV. The shapes of Figures 3 and 4 and the accompanying analysis are not affected by this consideration, and so it is ignored for expositional simplicity

the potential suppliers themselves. In this case the production process could itself be random. For example, the sellers could be hunters or gold miners. Alternatively, production could be such that each producer has a minimum level of sales below which it is uneconomical to produce at the current price. This level could vary among firms and be in the range where no firms find it optimal to limit production to an amount less than sales.

The formal analysis assumes that an exogenous producer supplies a flow of $S(\bar{P})$, $S' > 0$, units of the good to the market per unit of time. Each shop periodically receives shipments from this supplier in shipments of magnitude kS/M, where k is a positive constant and M denotes the number of shops.[5]

Denote the rate of sales at the open shops by R. The amount of time needed for a shop to sell one shipment of the good is therefore given by kS/RM. Assume that only sold-out shops receive shipments and that once a shop is sold out the arrival time of the next shipment is given by a random exponential process with mean $1/h$. On average therefore the amount of time between two shipments at any store is equal to $kS/RM + 1/h$. Each shop therefore receives an average flow of the good given by

$$(12)\qquad \frac{\frac{kS}{M}}{\frac{1}{h}+\frac{kS}{RM}} = \frac{\text{Shipment Size}}{\text{Average Time Between Shipments}}$$

For the inventory process to be consistent with an average flow supply of S to the market, it is required that the average flow to each shop equal S/M. This condition requires that h satisfy

$$(13)\qquad \frac{1}{h} = k\left(1-\frac{S}{RM}\right)$$

The average percentage of time that a shop is stocked will be given by the amount of time needed to sell a shipment as a fraction of the average time between shipments. In a market with many shops this fraction will also correspond to the average percentage of shops which are stocked and open at any time. Therefore, this fraction is given by

$$(14)\qquad \beta = \frac{\frac{kS}{RM}}{\frac{1}{h}+\frac{kS}{RM}} = \frac{S(\bar{P})}{RM}$$

The analysis of the inventory process allows us to express the fraction of shops open in terms of the flow supply per shop and the rate of sales at each shop.

III. The Excess Demand Steady State

We can now consider a steady-state situation. The analysis of the inventory process showed that the fraction of open shops is given by

$$(15)\qquad \beta = \frac{S(\bar{P})}{RM}$$

However, what determines the rate of sales at the open shops? Each period of time N new demanders enter the market. Upon entering the market each demander stays an average of $E(s)$ periods in the market before leaving. Therefore, on average, there will be $NE(s)$ demanders in the market at any time. Assuming that these demanders are distributed evenly over all the shops in each period of time, each shop will be visited by $NE(s)/M$ demanders per unit of time. With a prevailing price of $\bar{P}$, open shops will experience a rate of sales of

$$(16)\qquad R = \frac{D(\bar{P})E(s)}{M}$$

Since the expected duration of search depends upon the fraction of shops open, for consistency the equilibrium value of β must satisfy equation (16). However, in the previous section we noted that β must also satisfy equation (15). Combining equations (15) and (16) we find that the equilibrium fraction of shops open, β^*, must satisfy

[5]The analysis assumes that the shops play a passive role in their inventory strategy in order to focus more clearly on the effects of demander behavior. A more complete analysis would involve shops choosing their inventory policy so as to maximize expected profits = expected sales − inventory holding costs.

(17) $$\beta^* = \frac{S(\bar{P})}{D(\bar{P})E(s)}$$

Combining equation (17) with equation (10), we find that this value of β^* implies that

(18) $$S(\bar{P}) = q(\beta^*)\,D(\bar{P})$$

That is, the equilibrium value β^* has the property of equating the quantity purchased with the quantity supplied. This result should not be surprising, however, since it merely states that in equilibrium the inflow of goods to the market equals the flow of goods taken out of the market by the successful demanders. If β were less than β^*, the quantity purchased would be less than the quantity supplied. This would imply a net inflow of goods to the market. Over time more shops would become stocked and β would rise. If β were greater than β^*, the quantity purchased would be greater than the quantity supplied and the net outflow of the good would tend to cause a greater fraction of the stores to be sold out. Therefore β would fall. Figure 2 depicts the resulting steady-state solution.

The above analysis assumes that the rate of sales for each seller is less than some capacity rate. Otherwise, it would be necessary to consider the formation and length of queues. The present model suggests the following explanation for the formation of queues. If the above described steady state continued for a long time some potential sellers might become permanently inactive. This information might become known to the buyers so that the effective number of potential sellers might tend to shrink over time. However, the fraction of remaining potential suppliers that would have to be active at any time would remain constant for given values of $D(\bar{P})$ and N. Therefore the rate of sales for the remaining sellers when they were active would have to rise over time. Presumably if this rate rose enough it would eventually surpass the capacity rate and queues would form at the active sellers.

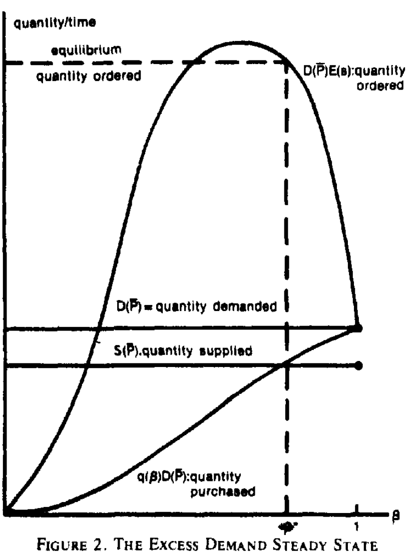


FIGURE 2. THE EXCESS DEMAND STEADY STATE

IV. Price and Availability Dynamics

Until now, the analysis has assumed that the price of the good has been fixed at $\bar{P}$. In the face of continuing excess demand, intuition dictates that in the absence of government control, the price would tend to rise. However, to date economic theory offers little in the way of a rigorous explanation of exactly how prices are set and how they change Proper analysis would explicitly identify the price setting agent and have him set prices in accordance to the maximization of some reasonable objective function.

The usual approach is the *ad hoc* assumption that the rate of price change is proportional to the level of excess demand. This assumption is intuitively pleasing and serves as a useful approximation for many purposes.

However, this approach is not the only plausible possibility. Suppose the individual sellers set prices. In the context of the model presented above, the individual sellers are capable of gauging the level of supply by observing the size and frequency of their deliveries. However, they are unable to directly observe the level of demand. The only indicator of demand available to the sellers is their rate of sales when they are stocked, or the number of potential customers

when sold out. That is, sellers are able to observe the quantity ordered, but not the quantity demanded. Therefore, it is reasonable to assume that the sellers adjust the price to equate the quantity ordered with the quantity supplied. In order to consider the effects of such a possibility, assume a price adjustment mechanism of the form

$$DP = \lambda_1 (D(P) E(s) - S(P)) \tag{19}$$

When the market is clearing, or when there is excess supply, the good will be readily available and may be purchased on the first search attempt. In such a situation the expected search time is one and equation (19) reduces to the more familiar form

$$DP = \lambda_1 (D(P) - S(P)) \tag{20}$$

However, during periods of excess demand, the expected duration of search may differ from one as some stores will generally be sold out, and so the quantity ordered will differ from the quantity demanded.

It is useful now to plot out those combinations of P and β which are consistent with price stability, in the context of equation (19). The equilibrium condition implied by equation (19) is that the quantity ordered equals the quantity supplied. That is,

$$E(s)D(P) = S(P) \tag{21}$$

Total differentiation of equation (21) shows that

$$\left.\frac{dP}{d\beta}\right|_{DP=0} = \frac{D(P)}{\frac{dS}{dP} - E(s)\frac{dD}{dP}} \cdot \frac{dE(s)}{d\beta} \tag{22}$$

The sign of expression (22) is the same as the sign of $dE(s)/d\beta$. It will be positive for low values of β and negative for high values of β.

Figure 3 presents a typical example of $DP = 0$ locus, with P^* indicating the price which equates supply and demand, and $\underline{P}$ indicating the price at which supply drops to zero. At those low values of β for which $E(s)$ is less than one, the quantity ordered is less than the quantity demanded. Therefore the price consistent with price stability must be below P^* in

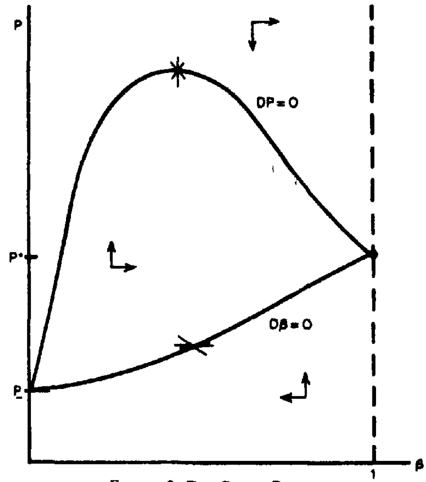


FIGURE 3. THE PHASE PLANE

order to lower supply relative to demand and equate the quantity supplied with the quantity ordered.

At higher values of β, the expected duration of search is greater than one and so the quantity ordered is greater than the quantity demanded. In this situation the price consistent with price stability must be greater than the market clearing price in order to increase supply relative to demand to equate the quantity supplied with the quantity ordered.

As was noted in Section III, β the fraction of shops open adjusts in such a way as to equate the quantity purchased with the quantity supplied. Assume that this adjustment may be characterized by

$$D\beta = \lambda_2 (S(P) - q(\beta)D(P)) \tag{23}$$

The equilibrium of this differential equation is given by

$$S(P) = q(\beta)D(P) \tag{24}$$

Differentiation of equation (24) with respect to P and β shows that

$$\left.\frac{dP}{d\beta}\right|_{D\beta=0} = \frac{D\frac{dq}{d\beta}}{\frac{dS}{dP} - q\frac{dD}{dP}} > 0 \tag{25}$$

and so the locus of values of P and β which satisfy equation (24) is upward sloping. Figure 3 also depicts this locus. Note that the two loci intersect at $\beta = 0$ and $\beta = 1$.

We are now ready to analyze the dynamics generated by the hypothesized adjustments of price and availability. The combined dynamic system is summarized by

$$DP = \lambda_1 (E(s)D(P) - S(P)) \tag{26}$$

$$D\beta = \lambda_2 (S(P) - q(\beta)D(P)) \tag{27}$$

Figure 3 therefore depicts the phase plane of this system. All points with the property

$$S(P) = D(P)E(s) = D(P)q(\beta) \tag{28}$$

characterize equilibria of this system. That is, equilibria are characterized by the equality of the quantities supplied, ordered, and purchased Two such equilibria exist. One is at $P = \underline{P}$, $\beta = 0$. For this solution the quantities supplied, ordered, and purchased are all zero. Analysis can show this equilibrium to be unstable.

In the other solution, the quantities supplied, ordered, and purchased are all equal to the quantity demanded. This corresponds to the conventional notion of equilibrium in a single market and is always stable. Here $P = P^*$ and $\beta = 1$.

We are now ready to trace out the time path the market will follow in adjusting to a situation of excess demand. Assume the market is initially in equilibrium with equality of the quantities supplied and demanded. Now assume some exogenous shock in supply or demand such that the prevailing price generates excess demand. This situation is depicted in Figure 4 as point A.

Since all stores are initially stocked, the quantities ordered and purchased will both equal the quantity demanded. With demand exceeding supply the quantity ordered will therefore also exceed the quantity supplied, and so the price will begin to rise. In addition, because the quantity purchased also exceeds the quantity supplied, there is a net outflow of the good from the market. This outflow soon causes some stores to become sold out, and so β begins to fall.

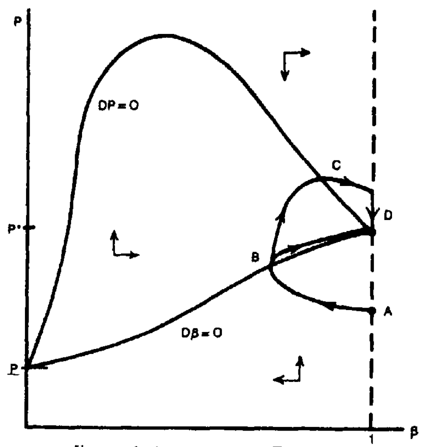


FIGURE 4 ADJUSTMENT TO EQUILIBRIUM

The rise in price will increase supply and reduce demand. This in conjunction with the fall in q induced by the fall in β will tend to equate the quantity purchased with the quantity supplied. This equality will be satisfied at point B. However, the quantity ordered will still exceed the quantity supplied, and so the price continues to rise.

This continued rise in price will diminish demand and increase supply so that now the quantity purchased will fall below the quantity supplied. This will translate into increased availability. That is, β will begin to rise. If the price adjusts slowly in comparison to β (the convergence of the stochastic inventory process) the adjustment will follow a path close to the $D\beta = 0$ locus. In this case the price will asymptotically approach P^* without overshooting. Such an adjustment path is sketched in Figure 4 as the lower path to point D.

On the other hand, if the price adjusts quickly in comparison to the inventory process, the price may rise above P^*. Above P^*, the quantity demanded will be less than the quantity supplied. However, because many demanders are making more than one order, the quantity ordered can still exceed the quantity supplied.

Therefore, the price continues to rise even though supply exceeds demand. Finally, the price will rise enough to equate the quantities ordered and supplied at point *C*.

Throughout this process, β has continued to rise as the quantity purchased has been less than the quantity supplied. The continuing rise in β will enhance the probability of a short search and so the quantity ordered will drop below the quantity supplied and so the price will begin to fall. Eventually, all the stores will be stocked. At this point the quantity ordered will equal the quantity demanded and the remaining adjustment to point *D* will be achieved through a fall in price proportional to the extent of excess supply.

REFERENCES

H. I. Grossman, "The Nature of Quantities in Market Disequilibrium," *Amer. Econ. Rev.*, June 1974, *64*, 509–14.

Bent Hansen, *A Study in the Theory of Inflation*, New York 1951.

Firm-Specific Evidence on Racial Wage Differentials and Workforce Segregation

By ROBERT HIGGS*

Two problems, one theoretical and the other empirical, currently obstruct the analysis of racial discrimination in the labor market. The theoretical problem arises from the profusion of models advanced to explain the data. These models rest on a mutually inconsistent variety of assumptions and, consequently, generate a mutually inconsistent variety of implications. The economist can now obtain a theoretical justification for any possible combination of racial wage differentials and workforce segregation: some models predict only the former, others only the latter, still others both or neither. Some predict a certain outcome only during the transition to an equilibrium, while others predict the same outcome as sustainable in equilibrium. (See the many models surveyed by Kenneth Arrow (1972a, b, 1973) and by Joseph Stiglitz.)

The existing models are similar insofar as each is a characterization of how firms behave in the labor market. Empirical testing of such models therefore requires information about the way firms hire and pay employees of different races. Instead—and here the empirical problem arises—data on incomes and occupations, often aggregative census data, have generally provided the raw materials for the econometrician's mill (see for example Gary Becker, Finis Welch, James Gwartney, Barry Chiswick, R. I. Mount and R. E. Bennett). Acceptance of these proxy variables has led to an apparent consensus that firms in this country, in fact, pay blacks less than identical whites (see Arrow 1973, p. 10, for such an inference). Well-documented racial differentials in incomes and occupational distributions have apparently been accepted as evidence of racial discrimination by firms in the payment of wages.[1] Yet this acceptance requires both a logical leap and an empirically undemonstrated assumption. Stiglitz has observed that "more empirical evidence is required to select among alternative hypotheses" (p. 295). He might have added that this selection will remain impossible unless the empirical evidence obtained is firm-specific.

Contemporary data are ill-suited for tests of the existing models. Available data generally provide information about the wages—but more often about earnings or incomes—of workers employed by an indeterminate number of unidentifiable employers. Even if one could somehow circumvent this difficulty, a far graver problem would remain. The existing models characterize the market behavior of firms in the absence of legal constraints on discriminatory acts: except for the market repercussions they may experience as a result of their discriminatory behavior, firms are free to discriminate if they choose to do so. Such freedom is obviously not available to firms at the present time in the United States. State fair employment laws, federal contract restrictions, and a variety of similar legal constraints, when enforced, directly penalize firms for discriminatory acts in the labor market. This does not mean, of course, that firms never discriminate. But it does mean that employers who practice discrimination have an incentive to conceal that fact. Even if one could obtain reliable firm-specific observations, they would be generated in a context incompatible with that assumed by the existing models and would therefore not constitute appropriate evidence for a test of those models.

One can obtain proper evidence, however,

*University of Washington. I am grateful for helpful comments received from many colleagues at Washington and from Stanley Engerman, Donald McCloskey, Edward Meeker, and Alan Olmstead

[1]Ray Marshall, p 862, explicitly rejects this inference.

from an earlier period of American history. The present paper presents data compiled from almost 20,000 firm-specific observations of wages paid to blacks and whites in various narrowly defined nonagricultural occupations in Virginia in 1900 and 1909.[2] The labor markets studied can be presumed to contain white employers and employees interested in racial discrimination. No laws obstructed racial wage discrimination, and indeed racial discrimination in a variety of other forms was not only blatantly practiced but publicly applauded by the leading lights of that time and place. Besides presenting evidence on racial wage differentials, this paper offers evidence on the extent of workforce segregation in a number of occupations, and examines the interrelations of racial wage differentials with workforce segregation. Along the way, it provides some lessons about the grave potential pitfalls of inferences about average firm behavior drawn from evidence on the average wages received by groups of workers, lessons made quite clear when the two kinds of data are set side by side.

I. The Data

The evidence to be examined comes from the *Fourth Annual Report* and the *Thirteenth Annual Report* of the Virginia Bureau of Labor and Industrial Statistics. These reports provide data for the calendar years 1900 and 1909, both of which may be regarded as "normal" years, neither boom nor bust in the business cycle.[3] The bureau collected this information as part of its ongoing effort to describe industrial and labor conditions in the state. Thousands of firms responded to its requests for information on value of product, wages paid, hours of work, capital invested, industrial accidents, and other subjects. Although the firms are identified only by number, and therefore one cannot ascertain their locations, it seems likely that they were distributed throughout the state.[4]

The bureau did *not* collect these data for the purpose of facilitating study of racial wage differentials or workforce segregation. Yet for a number of occupations, data were reported and compiled separately for black and white workers. These racial distinctions were "natural" under the circumstances of the time and place. Racism was then rampant, and the state's recordkeeping did not escape it; even tax returns were recorded separately for the two races. That racial distinctions were made in an incidental, almost unconscious, manner in these investigations may actually augment their value to the student of racial differences. In any event, there is no reason to suppose that deliberate falsification or sampling biases distort the revealed racial differentials in wages or the patterns of workforce segregation.

From the two reports, I have selected for study all those occupations with wage data for at least 25 workers of each race. This selection criterion yields eleven occupations for 1900 and fourteen for 1909. My unit of observation is the "contract," by which I mean the daily wage rates paid by a specific firm to workers of each race employed in a particular occupation. Firms reported this information in the form, say, 12 white carpenters at $2.00 per day, 3 black carpenters at $1.75 per day. Because some firms hired workers for more than one sample occupation, the number of contracts in the sample exceeds the number of firms. The samples cover 290 firms making 490 contracts in 1900, and 636 firms making 1,595 contracts in 1909; they include 5,292 workers (2,247 white) in 1900 and 13,995 (6,937 white) in 1909. In view of the

[2] For evidence that racial wage differentials in *agricultural* employment at the turn of the twentieth century were either very small or nonexistent, see the author (1972, 1975, 1977, ch. 4).

[3] The estimated percentage of the U.S. civilian labor force unemployed was 5.0 in 1900 and 5.1 in 1909 (see Stanley Lebergott, p. 512).

[4] In its only explicit statement about coverage, the report for 1900 says that "reports were received from nearly all the contractors in the several building trades . . . [yielding] almost a complete compilation . . ." (p. 1). No doubt coverage was less complete in the other industries surveyed, as one can confirm by comparing the numbers in the survey with the numbers reported in the federal censuses of occupations.

TABLE 1—DISTRIBUTION OF SAMPLE CONTRACTS BY RACIAL COMPOSITION OF OCCUPATIONAL WORKFORCE

Occupation	Contracts for hire of: Whites only	Blacks only	Both races	Total contracts
1900				
Carpenters	78	1	15	94
Lathers	7	3	6	16
Brickmakers	4	4	4	12
Tannery beamsmen	5	3	8	16
Plumbing laborers	7	33	14	54
Sawmill teamsters	42	32	11	85
Loggers	33	30	6	69
Sawmill laborers	36	42	23	101
Bark grinders	4	4	7	15
Tannery yardmen	4	2	8	14
Tannery misc help	4	2	8	14
Totals	224	156	110	490
1909				
Brickmakers	5	19	13	37
Brickworks kilnmen	12	12	6	30
Bricklayers	69	3	8	80
Carpenters	204	2	15	221
Sawmill engineers	136	38	13	187
Sawyers	153	14	8	175
Brickworks laborers	7	11	12	30
Bricklayers' helpers	7	7	1	15
Gen contr. laborers	41	65	40	146
Plumbing laborers	11	87	16	114
Sawmill laborers	73	70	61	204
Loggers	57	39	22	118
Sawmill misc help	38	22	24	84
Sawmill teamsters	63	63	28	154
Totals	876	452	267	1,595

variety of occupations and the large number of firms, contracts, and workers represented, the sample provides a firm basis for generalizations.

Table 1 shows the distribution of sample contracts by racial composition of the occupational workforces; Table 2 shows the distribution of sample workers by race, occupation, and racial composition of the occupational workforce to which they belong.

Moreover, these data are probably representative of conditions not only in Virginia but throughout the South. The sample occupations were commonly pursued throughout that region. Nothing in Virginia's economic development, industrial structure, labor market conditions, or prevailing patterns of race relations set it markedly apart from other southern states. Inasmuch as nine-tenths of the black population still lived in the South at the beginning of the twentieth century, evidence from Virginia furnishes a window for viewing the labor market conditions faced by the overwhelming majority of the black workforce. Further, evidence for the first decade of the twentieth century is probably representative of conditions as they existed for several decades before and after that period. One can with justification consider the data examined here as more than a case study: these data probably provide a sample representative of conditions over a much wider geographical and temporal domain.

II. Racial Wage Differentials

Did a firm pay less to its black than to its white laborers? To answer this question, of course, one must examine only firms that actually hired both races in a given occupation. Table 3 shows the number of sample contracts covering integrated occupational workforces, distributed according

TABLE 2—DISTRIBUTION OF SAMPLE WORKERS BY RACE, OCCUPATION, AND RACIAL COMPOSITION OF WORKFORCE

Occupation	Whites A	Whites B	Blacks A	Blacks B	Total
1900					
Carpenters	615	179	3	43	840
Lathers	12	27	10	16	65
Brickmakers	14	39	43	93	189
Tannery beamsmen	29	84	14	101	228
Plumbing laborers	18	20	79	24	141
Sawmill teamsters	118	34	179	34	365
Loggers	150	81	150	367	748
Sawmill laborers	202	372	541	1,234	2,349
Bark grinders	8	25	8	27	68
Tannery yardmen	39	77	4	38	158
Tannery misc. help	27	77	4	33	141
Totals	1,232	1,015	1,035	2,010	5,292
1909					
Brickmakers	34	162	276	344	816
Brickworks kilnmen	40	7	60	24	131
Bricklayers	288	45	4	24	361
Carpenters	1,892	120	8	34	2,054
Sawmill engineers	182	20	44	21	267
Sawyers	202	14	21	16	253
Brickworks laborers	148	56	124	159	487
Bricklayers' helpers	32	7	38	7	84
Gen contr laborers	159	494	520	479	1,652
Plumbing laborers	32	45	228	43	348
Sawmill laborers	993	537	1,468	1,388	4,386
Loggers	338	171	228	333	1,070
Sawmill misc help	339	214	290	492	1,335
Sawmill teamsters	243	123	205	180	751
Totals	4,922	2,015	3,514	3,544	13,995

Note: Columns A denote "In segregated work forces", columns B denote "In integrated work forces."

to whether they paid whites more, both races the same, or blacks more. (Throughout this paper, unless otherwise stated, the words "less," "same," and "more" are used in a literal, not a statistical, sense.) In 1900, 36 percent of the contracts paid whites more, 61 percent paid both races the same, and 3 percent paid blacks more; in 1909 the respective percentages were 38, 57, and 5. Clearly, from the standpoint of firm behavior, equal payment within a given occupation was the most common practice. In the absence of information on the skills, experience, and other productivity attributes of the workers within a given occupation, one cannot say whether the instances where a firm paid more to whites indicate discriminatory behavior.

One can gain further information on this question, however, by distinguishing occupations according to their prevailing requirements of skills and experience. Under the maintained hypothesis that the observed racial wage differentials reflected differentials in worker productivity in a situation where the whites were better trained and more experienced on the average, one could predict that racial differentials would be more commonly observed in skilled occupations than in unskilled occupations. In the latter there is simply relatively little potential for productivity-associated heterogeneity in the workforce. When the contracts are distributed according to the skill levels of the occupations, the results are as shown below.[5]

[5]The distinction of skilled (including semiskilled) versus unskilled rests partly on job descriptions but mainly on prevailing wage levels. See Table 3 for the classification. Clearly, all the occupations classified as unskilled require neither much training nor long experience. As always, however, imposing a discrete categorization on a continuous variable is a rough procedure at best.

TABLE 3—DISTRIBUTION OF SAMPLE CONTRACTS FOR HIRE OF INTEGRATED OCCUPATIONAL WORKFORCES BY RACIAL DIFFERENCE IN WAGES PAID

Occupation	Number of contracts paying: Whites more	Both same	Blacks more	Total contracts
1900				
Carpenters[a]	12	1	2	15
Lathers[a]	4	2	0	6
Brickmakers[a]	4	0	0	4
Tannery beamsmen[a]	0	7	1	8
Plumbing laborers	3	11	0	14
Sawmill teamsters	2	9	0	11
Loggers	2	4	0	6
Sawmill laborers	8	15	0	23
Bark grinders	1	6	0	7
Tannery yardmen	2	6	0	8
Tannery misc help	2	6	0	8
Totals	40	67	3	110
1909				
Brickmakers[a]	8	5	0	13
Brickworks kilnmen[a]	4	2	0	6
Bricklayers[a]	6	2	0	8
Carpenters[a]	15	0	0	15
Sawmill engineers[a]	5	8	0	13
Sawyers[a]	3	4	1	8
Brickworks laborers	4	6	2	12
Bricklayer helpers	1	0	0	1
General contract laborers	14	19	7	40
Plumbing laborers	2	13	1	16
Sawmill laborers	14	44	1	59
Loggers	5	17	0	22
Sawmill misc help	11	13	0	24
Sawmill teamsters	8	16	0	24
Totals	100	149	12	261

[a]Occupation classified as skilled or semiskilled

	Occupations skilled	unskilled
In 1900:		
Contracts paying blacks same or more	13	57
Contracts paying whites more	20	20
In 1909:		
Contracts paying blacks same or more	22	139
Contracts paying whites more	41	59

Chi-square test statistics for these cross classifications are 11.98 for 1900 and 25.62 for 1909, both highly significant. These tests for association warrant the conclusion that skill levels and the payment of racial wage differentials were not independent. Obviously, the sample firms were much more likely to pay whites more than blacks in the skilled occupations. While this result does not justify rejection of the hypothesis that at least some firms paid purely discriminatory premiums to whites, the data are also consistent with the hypothesis that racial wage differentials reflected productivity differentials; and the latter hypothesis explains something that the former does not, namely, the much greater frequency of differentially higher wages paid to whites in the skilled occupations, where the whites ranked higher in terms of skills and experience on the average.[6]

III. Statistical Asymmetries

To make inferences about the behavior of the

[6]For evidence that white workers were more productive on the average, see the author (1977, ch 4) and Richard Freeman (1972a, b).

average firm based on evidence about the average wage received by a group of workers is to risk faulty interpretation. Computing the means and standard errors of wages *received* by workers in the sample occupations, one discovers that in 1900 the mean wage was significantly greater for the whites in seven of the eleven occupations (the four tannery occupations being the exceptions); in 1909, whites received significantly more in twelve of the fourteen sample occupations (brickworks kilnmen and general contract laborers being the exceptions). Yet the firm-specific observations compiled in Table 3 show that in 1900 contracts paying blacks the same as or more than whites outnumbered contracts paying more to the whites in eight of the eleven occupations; and in 1909 the corresponding figure was nine out of fourteen. Clearly, looking at the question of firm behavior—which is what the existing models of labor market discrimination are about—from the workers' end of the transaction can be seriously misleading if conditions vary across firms, which empirically they always do.

Sample laborers employed by plumbers in 1909 provide a concrete and illuminating example. In that occupation the 77 white laborers earned a mean wage of $1.56 per day, the 271 blacks $1.39. The difference is highly significant statistically. But of the 16 firms hiring integrated workforces for this occupation, only 2 paid the whites more. In this case, which is similar to many others in the sample, an examination of wages received, without any knowledge of how workers are distributed across firms paying quite different wages, conveys not just limited information; rather, it conveys a completely misleading impression. The average white wage was so much higher than the average black wage in this occupation because a very large proportion of the whites (20 workers) worked for a firm that paid a very atypical $2.00 per day. This same firm hired 10 blacks at $1.50 per day, which, although less than what the firm paid its white workers, was still a wage surpassed by only 3 white workers in other firms.

Sawmill laborers in 1909 illustrate a different kind of statistical asymmetry. In this occupation the 1,530 white workers received a mean wage of $1.22 per day, the 2,856 blacks $1.15. The difference is highly significant statistically. As Table 3 shows, however, 44 of the 59 firms hiring both races for this occupation paid them the same wage. Looking closer, one finds that the white workers in integrated workforces received a mean wage of $1.16, the black workers $1.17, the difference being statistically insignificant. In this case, which resembles some others in the sample, the whites earned more in the overall sample because the 993 whites working in segregated workforces got a mean wage of $1.25, while the blacks in segregated workforces got a mean wage of $1.14.

IV. Workforce Segregation

Within the sample contracts, occupational workforce segregation was overwhelmingly the rule: in 1900, 46 percent of the contracts hired for a given job only whites, 32 percent only blacks, and 22 percent both races; in 1909 the respective percentages were 55, 28, and 17. Of the workers, in 1900, 55 percent of the whites and 34 percent of the blacks worked in segregated occupational workforces; in 1909 the respective percentages were 71 and 50.

The differences between the contract and worker distributions by segregation status imply that contracts for the hire of integrated workforces involved a larger average workforce, which is hardly surprising. If a firm hires only one worker for a given job, as many firms in the sample did, its occupational workforce is necessarily segregated. Even if it hires two or three, the chances of getting a segregated workforce are quite high if the workers are drawn randomly from a working population in which the two races are very disproportionately represented. If a proportion β of the potential workers are white, then the probability of a firm's randomly hiring a workforce of n workers, all of them white, is β^n. Given that labor markets were geographically limited and that the two races were distributed quite unequally over the state—blacks predominated in the southeastern and whites in

the western parts of the state—it was almost inevitable that many workforces would be segregated even if neither employers nor employees valued segregation per se. But as the size of the workforce increased, a random process was increasingly unlikely to generate a segregated workforce (β^n approaches zero as n grows larger), and large segregated workforces are strong evidence that deliberate action had been taken to attain that result.

To test how much of the workforce segregation in the sample can be statistically explained by the size of the workforce, a regression relation has been fitted to the cross-sectional observations of occupations.[7] This regression also includes a dummy variable to control for skill levels in the occupations, as inspection of the data suggests that skills and segregation may not have been independent across occupations. Letting I = the percentage of contracts with integrated workforces, W = the mean size of the workforce, and S = a dummy variable equal to unity for skilled occupations and zero otherwise, one obtains the following ordinary least squares equations (standard errors in parentheses):

$$1900: \quad I = 32.41 + \underset{(1.0030)}{0.0763}\, W + \underset{(12.3338)}{1.0143}\, S$$

$$N = 11, SE = 19.58, R^2 = 0.002$$

$$1909: \quad I = 8.95 + \underset{(0.2600)}{1.2830}\, W - \underset{(3.5835)}{4.1675}\, S$$

$$N = 14, SE = 6.38, R^2 = 0.74$$

For 1900 the equation completely fails to fit the observed data, but for 1909 it accounts for three-fourths of the variance across occupational observations. This difference need not surprise anyone; after all, only five occupations are common to the two samples, and so the equations apply to quite differently composed sets of observations as well as different dates. The results for 1909 indicate that the percentage of contracts with integrated workforces was significantly related to the average size of the workforce across occupations. There is also an indication that more skilled occupations were less integrated on the average, but the significance of this partial relation is too low to warrant its acceptance at conventional test levels.

Inasmuch as the 1900 sample is much smaller than the 1909 sample, I am inclined to take the insignificant test results for the former year somewhat less seriously than the significant results for the latter year. To the extent that one accepts the relation between the prevalence of integrated workforces and the average size of the workforce, one also diminishes the extent to which workforce segregation can be viewed as the result of discriminating behavior by white employers or employees. The hypothesis of "hostile" segregation, of course, makes no prediction about the relation of segregation to workforce size (except the trivial prediction discussed in the next paragraph). It therefore fails to account for something that the hypothesis of random selection explains fairly well, at least for the sample of 1909.

Of course, any explanation of the extent of integration must recognize that firms which hire only a single worker for a given occupation necessarily create a segregated occupational workforce. To determine how much of the variation in the percentage of contracts with integrated workforces can be statistically explained by this one-man workforce effect, one can correlate I with the percentage of firms hiring only a single worker for a given occupation. Across the occupations of 1909, the result is: $r = -0.53$; hence, $r^2 = 0.28$. But while the percentage of firms with one-man occupational workforces statistically accounts for 28 percent of the variance in I, the average size of the workforce accounts for 71 percent ($r = 0.84$). This difference confirms that the significance of W, as shown in the regression, is not spurious. The relation between integration and the size of the workforce was genuine, at least in 1909, and obtained among firms with more than one worker in their occupational workforces as well as among all sample firms.

[7] Data for the regression are computed from basic data shown in Tables 1 and 2.

V. Interrelations of Segregation and Racial Wage Differentials

Were workforce segregation and racial wage

differentials related? To answer this and other questions about the sample workers, I have constructed Table 4, which shows several varieties of wage ratios. (In the table, *BSEG* and *BINT* denote segregated and integrated blacks, *WSEG* and *WINT* segregated and integrated whites, respectively.) These results are not easily summarized, but in a rough way they indicate that, in terms of *average wages received,* the groups tended toward an ascending order: segregated blacks, integrated blacks, integrated whites, segregated whites. However, the first two groups are very close, as are the last two. Perhaps a firmer conclusion is simply that the blacks generally averaged less than the whites in a given occupation. Within an occupation, members of a particular race tended to receive about the same wage regardless of whether they worked in integrated or segregated workforces. (Table 4 also makes a contribution by establishing the orders of magnitude for the racial wage differentials that did exist.)

VI. Conclusions

Ever since emancipation the main thrust of racial economic discrimination in America has been not the payment of lower wages to blacks than to productively identical whites, but rather a variety of measures that have had the effect of keeping the productivity of black labor low. As a consequence, relatively few blacks have been qualified to perform the higher-paying, higher-status jobs. To keep the blacks "in their place," the whites for a century discriminated strongly against them in the public sector through the actions of police, law courts, and public schools,

TABLE 4—RATIOS OF DAILY WAGES RECEIVED BY OCCUPATION, RACE, AND RACIAL COMPOSITION OF OCCUPATIONAL WORKFORCE

Occupation	*BSEG*/*WSEG*	*WSEG*/*WINT*	*BSEG*/*BINT*	*BINT*/*WINT*
1900				
Carpenters[a]	nc	0.95	nc	0.65
Lathers[a]	0.72	1.03	0.78	0.95
Brickmakers[a]	0.80	1.08	1.23	0.70
Tannery Beamsmen[a]	0.84	0.97	0.78	1.05
Plumbing laborers	0.89	1.06	1.05	0.90
Sawmill teamsters	0.92	1.04	1.13	0.85
Loggers	0.78	0.93	0.83	0.88
Sawmill laborers	0.77	0.90	0.77	0.90
Bark grinders	nc	nc	nc	1.01
Tannery yardmen	nc	1.01	nc	1.03
Tannery misc. help	nc	1.01	nc	1.00
1909				
Brickmakers[a]	1.15	0.67	0.93	0.83
Brickworks kilnmen[a]	1.02	0.73	1.03	0.72
Bricklayers[a]	nc	1.11	nc	0.67
Carpenters[a]	nc	0.99	nc	0.61
Sawmill engineers[a]	0.76	1.01	0.93	0.82
Sawyers[a]	0.75	1.00	0.80	0.95
Brickworks laborers	0.94	1.04	1.02	0.96
Bricklayer helpers	1.00	nc	nc	nc
General contract laborers	0.97	1.13	0.98	1.12
Plumbing laborers	0.99	0.83	0.97	0.85
Sawmill laborers	0.91	1.08	0.97	1.01
Loggers	0.86	1.17	1.06	0.96
Sawmill misc. help	0.81	1.25	1.08	0.94
Sawmill teamsters	0.88	1.06	0.99	0.94

Note: nc denotes not computed because of small sample size (i.e., either numerator or denominator of the ratio represents less than 10 workers). Column headings are defined in the text.

[a]Occupation classified as skilled or semiskilled

thereby attenuating black property rights and civil liberties, depriving blacks of access to high quality education and training, and reducing their mobility. In their acts of discrimination, the whites have generally sought not to avoid physical proximity to blacks but rather to insure that the blacks would "naturally" occupy a subordinate social and economic position.[8]

Perhaps most economists with some knowledge of black history would accept these propositions; yet many persist in believing that the payment of lower wages for homogeneous labor services supplied by blacks has also played an important role. Evidence presented above, however, indicates that this belief is ill-founded. In 1900, of 77 sample contracts for the hire of integrated unskilled workforces, only 20 paid more to whites; in 1909, only 59 of 198 such contracts paid more to whites. And in the absence of more detailed knowledge about worker characteristics—age, literacy, absenteeism, and so forth—it is impossible to say whether that minority of contracts paying less to blacks discriminated in the strict sense of paying less for homogeneous labor services supplied by blacks. Workforce segregation was much more common than the payment of racial wage differentials within integrated workforces. But much of the observed segregation probably sprang from the conjunction of small workforces with lopsided racial distributions of potential workers rather than from deliberate choice in the service of racial hostility. In any event, workers on the average got about the same wage within a given occupation whether they worked in integrated or segregated occupational workforces.

That white workers in integrated workforces did not systematically receive a higher wage than white workers in segregated workforces refutes an important prediction of most models of discrimination (see Welch, pp. 228–231; Arrow 1972b, p. 197; Chiswick, p. 1346). Further, the data are inconsistent with those models that predict complete segregation (see Arrow 1972a, p. 92; 1972b, p. 197) and with those that predict integration will increase the average wage of each group (Welch, pp. 228–29). While the data do not warrant the sweeping conclusion that discriminating firms could not survive, they strongly suggest that competitive pressures resulted in equal payment for equal labor services in most cases. In addition, the association of racial wage differentials with skilled employment suggests that racial productivity differentials underlay at least some part of the observed racial wage differentials. The data examined here are insufficiently detailed to permit further testing of this hypothesis, but on the basis of my related research (summarized in my 1977 book) I am willing to conjecture that once productivity-associated racial differences have been controlled, only a small minority of firms will be found to have discriminated in the sense of paying less for homogeneous labor services supplied by blacks.

To account for the observed racial differentials in incomes and occupations, economic analysis must come seriously to grips with the complex of historical, cultural, and nonmarket influences which, as Ray Marshall has recently expressed it, "makes it less likely that black and white workers will be homogeneous substitutes" (p. 861). This will require considerable extension of our models, to give them an explicit temporal dimension and to incorporate more variables presently regarded as noneconomic. It will also require more careful attention to the role of collective, as opposed to purely individualistic, discrimination.[9] Such requirements will not be easily satisfied, of course; but to continue along presently defined lines is to condemn ourselves to a sterile, if elegant, analysis and to limit severely our ability to understand and predict actual events.

[8] See the studies of white objectives by the historians G. M. Fredrickson and Lawrence Friedman.

[9] In the second edition of his pioneering book, Becker observes that "our ignorance of the scope and incidence of collective action against minorities is perhaps the most important remaining gap in the analysis of the economic position of minorities" (p. 8)—surely an understatement.

REFERENCES

K. J. Arrow, (1972a) "Models of Job Discrimination," in Anthony H. Pascal, ed., *Racial Discrimination in Economic Life,* Lexington 1972, 83–102.

———, (1972b) "Some Mathematical Models of Race Discrimination in the Labor Market," in Anthony H. Pascal, ed., *Racial Discrimination in Economic Life,* Lexington 1972, 187–203.

———, "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets,* Princeton 1973, 3–33.

Gary S. Becker, *The Economics of Discrimination,* Chicago 1957.

B. R. Chiswick, "Racial Discrimination in the Labor Market: A Test of Alternative Hypotheses," *J. Polit. Econ.,* Nov./Dec 1973, *81,* 1330–52.

G. M. Fredrickson, *The Black Image in the White Mind: The Debate on Afro-American Character and Destiny, 1817–1914,* New York 1971.

R. B. Freeman, (1972a) "Black-White Economic Differences: Why Did They Last So Long?," unpublished 1972.

———, (1972b) "Long Term Changes in Black Labor Market Status: A Preliminary Report," unpublished 1972.

Lawrence J. Friedman, *The White Savage: Racial Fantasies in the Postbellum South,* Englewood Cliffs 1970

J. Gwartney, "Discrimination and Income Differentials," *Amer. Econ. Rev.,* June 1970, *60,* 396–408.

R. Higgs, "Did Southern Farmers Discriminate?," *Agr. History,* Apr. 1972, *46,* 325–28.

———, "Did Southern Farmers Discriminate? —Interpretive Problems and Further Evidence," *Agr. History,* Apr. 1975, *49,* 445–47.

———, *Competition and Coercion: Blacks in the American Economy, 1865–1914,* New York 1977.

Stanley Lebergott, *Manpower in Economic Growth: The American Record since 1800,* New York 1964.

R. Marshall, "The Economics of Racial Discrimination: A Survey," *J. Econ. Lit.,* Sept. 1974, *12,* 849–71.

R. I. Mount and R. E. Bennett, "Economic and Social Factors in Income Inequality: Race and Sex Discrimination and Status as Elements in Wage Differentials," *Amer. J. Econ. Soc.,* Apr. 1975, *34,* 161–74.

J. E. Stiglitz, "Approaches to the Economics of Discrimination," *Amer. Econ. Rev. Proc.,* May 1973, *63,* 287–95.

F. Welch, "Labor-Market Discrimination: An Interpretation of Income Differences in the Rural South," *J. Polit. Econ.,* June 1967, *75,* 225–40.

Virginia Bureau of Labor and Industrial Statistics, *Fourth Annual Report,* Richmond 1901.

———, *Thirteenth Annual Report,* Richmond 1910.

A Note on Short-Run Asset Effects on Household Saving and Consumption

By FREDERIC S. MISHKIN*

In an interesting and valuable article in this *Review*, Irwin Friend and Charles Lieberman (F-L) use Federal Reserve cross-section data to test for the effects of capital gains on household saving. F-L state, "The bulk of our estimates, which are concentrated in the lower half of the .02 to .04 range, [i.e., a dollar of capital gains leads to increased consumption of two to four cents] are . . . modestly less than the figure implied by the latest version of the *MPS* [MIT-PENN-SSRC] model" (p. 625). F-L also state that "the *FRB* survey results suggest that the latest *MPS* estimate of the first year's effects of a change in the value of stock assets on saving and consumption is not far off from the true figure though it may be a little on the high side . . ." (p. 632).

Both of the above statements are based on an incorrect F-L calculation of .032 for the capital gains coefficient on consumption implied by the latest unpublished version of the *MPS* model. Instead of calculating the implied *MPS* marginal propensity to consume out of capital gains for the 1963 year (which would be comparable with their cross-section estimates), F-L calculate the capital gains effect on the annual rate of consumption in the last quarter of that year.[1] Since capital gains accrue fairly evenly over the year in question (1963), the capital gains effect calculated by F-L will exceed the capital gains effect on the full year's consumption. Using a similar allocation of capital gains to each quarter of 1963 as F-L (see fn. 23) and using the *MPS* consumption function and the *MPS* definition of quarterly stock market wealth, I derived the *MPS* implied coefficient of capital gains on consumption to be .013.[2] (See the Appendix.)

The .013 coefficient, which is consistent with the equations and definitions in the *MPS* model, is substantially smaller than the .032 F-L figure and leads to a different interpretation of their results. The *MPS* implied coefficient is no longer above most of the F-L estimates, but is rather smaller than all the capital gains coefficient estimates in their Table 1 (the smallest F-L estimate is .015) and is statistically significantly lower than many of the F-L estimates.[3]

There are several possible conclusions from a comparison of the corrected implied *MPS* capital gains coefficient and the F-L empirical estimates.

1) If the F-L cross-section results are felt to be more accurate than the *MPS* time-series estimates, the *MPS* wealth effects on consumption may be underestimated rather than modestly overestimated, as implied by F-L.

2) The higher F-L estimates of the effects of capital gains may indicate that the distributed lag on stock market wealth in the *MPS* consumption function is too long. A shorter distributed lag would imply that changes in stock

*Assistant professor, department of economics, University of Chicago. This comment was written while I was the recipient of a NSF Graduate Fellowship

[1]The .032 figure of F-L is only approximately equal to the *MPS* capital gains effect on the annual rate of consumption for the last quarter of that year, because F-L do not make use of an averaging procedure for the valuation of stock assets which is a feature of the *MPS* model. The implied *MPS* capital gains effect on the annual rate of consumption for the last quarter of that year would be slightly lower than the F-L .032 figure for this reason The Appendix describes the F-L calculation

[2]The implied coefficient in the last published version of the *MPS* model (found in Franco Modigliani) is by an analogous procedure calculated to be .018.

[3]The null hypothesis that the F-L capital gains coefficient is equal to the *MPS* .013 figure is rejected at the 5 percent level or higher for equations (2), (3), (4), (5), (10), (12) in F-L's Table 1: that is, in half of the cases. The t-statistics of this hypothesis test for these equations are respectively 2.54, 2.54, 2.60, 4.31, 2.69, 4.80. The critical t at the 5 percent level for a two-tailed test is 1.96, while the critical t at the 1 percent level is 2.58.

market wealth affect consumption faster, and hence the one-year capital gains coefficient would be higher.

3) The *MPS* estimate may be understated relative to the F-L estimates because F-L exclude from their capital gains measure capital gains on stock accruing in pension funds, while the *MPS* model does not. The exclusion of pension fund capital gains may well be appropriate because capital gains accruing in pension funds often do not directly affect the benefits that the household will receive from these funds. Whether this exclusion is appropriate is an empirical matter that requires further tests.[4]

One further point is worth making As a result of a failure to differentiate between stocks[5] and flows, many authors have assumed that a finding of small one-year capital gains effects is inconsistent with the results of the *MPS* consumption function The small .013 size of the *MPS* capital gains coefficient implies that earlier studies which find small one-year asset effects (see F-L references) are not necessarily inconsistent with the *MPS* results. This is especially true of studies by John Arena and Kul Bhatia where the long-run wealth effect is close to the *MPS* .054 figure and the capital gains effect is near the *MPS* .013 value.[6]

Because the magnitude of the stock market capital gains effect can be so critical in stabilization policy and the determination of aggregate demand, as is indicated by the *MPS* model, the issues raised by F-L are indeed important. The Friend and Lieberman paper lends strong support to the view that there are large wealth effects on aggregate demand through common stock capital gains. Indeed, their paper indicates that wealth effects may affect aggregate demand faster and be even more potent than is indicated by the *MPS* model.

[4]As a result of the structure of the *MPS* model, the inclusion or exclusion of capital gains in pension funds would make very little difference to the simulation characteristics and thus the policy implications of the *MPS* model

[5]The word "stocks" here refers to the accounting concept of stocks and does not refer to common stocks

[6]The Arena consumption function most comparable to the *MPS* consumption function gives a long-run wealth effect of 038 and a capital gains effect of .015, while Bhatia finds that the long-run wealth effect is .052 and the capital gains effect is .014

APPENDIX

Calculation of the MPS Implied Capital Gains Coefficient

The *MPS* consumption function states that the additional flow (annual rate) of consumption within the quarter from stock market capital gains effects is,

$$CC_t = \sum_{i=1} C_i V_{t-i+1} \tag{A1}$$

where CC_t = additional consumption at an annual rate for quarter t

V_t = average increased value of stocks —beginning of quarter t

The actual quarterly change in consumption is obtained by dividing equation (A1) through by four.

$$CQ_t = \frac{\sum_{i=1} C_i V_{t-i+1}}{4} = \sum_{i=1} \left(\frac{C_i}{4}\right) V_{t-i+1} \tag{A2}$$

where CQ_t = additional consumption in quarter t.

In the *MPS* model the beginning of quarter average value of stocks is defined as the average value of stocks over the t and t − 1 quarters. Thus,

$$V_t = \frac{X_t + X_{t-1}}{2} \tag{A3}$$

where X_t = average increased value of stocks in quarter t

Using equations (A2) and (A3) to calculate the additional consumption for each quarter as a result of capital gains accruing over the year, we have

$$CQ_1 = \frac{C_1}{4}\left[\frac{X_1 + 0}{2}\right] \tag{A4}$$

$$CQ_2 = \frac{C_1}{4}\left[\frac{X_2 + X_1}{2}\right] + \frac{C_2}{4}\left[\frac{X_1 + 0}{2}\right] \tag{A5}$$

$$CQ_3 = \frac{C_1}{4}\left[\frac{X_3 + X_2}{2}\right] + \frac{C_2}{4}\left[\frac{X_2 + X_1}{2}\right] + \frac{C_3}{4}\left[\frac{X_1 + 0}{2}\right] \tag{A6}$$

$$\text{(A7)} \quad CQ_4 = \frac{C_1}{4}\left[\frac{X_4 + X_3}{2}\right] + \frac{C_2}{4}\left[\frac{X_3 + X_2}{2}\right] + \frac{C_3}{4}\left[\frac{X_2 + X_1}{2}\right] + \frac{C_4}{4}\left[\frac{X_1 + 0}{2}\right]$$

The total effect of capital gains on consumption for the year is just the sum of the four quarterly effects.

$$\text{(A8)} \qquad CGE = \sum_{t=1}^{4} CQ_t$$

where CGE = total capital gains effect.

The average capital gains in each quarter of 1963 have been allocated using the Standard and Poor's composite stock index.[7] From January 1 to December 31 of 1963, the index rose from 62.69 to 75.02—an appreciation of 12.33 points. The quarterly average of the index for the four quarters of 1963 are respectively, 65.54, 69.67, 70.97, and 73.27. Subtracting 62.69 from the quarterly averages and dividing by 12.33, we get the average quarterly accrual per dollar of 1963 capital gains: i.e., $X_1 = .2311$, $X_2 = .5432$, $X_3 = .6715$, $X_4 = .8581$ for the first through fourth quarters, respectively. The *MPS* consumption function coefficients for the first four lag quarters of stock market wealth are $C_1 = .016421$, $C_2 = .012742$, $C_3 = .009540$, and $C_4 = .006788$. These values are used with equations (A4) through (A8) to derive the .013 figure found in the text of this comment. The implied capital gains coefficient for the last published version of the *MPS* model (in Modigliani) is calculated using the same procedure with the lag coefficients found in Modigliani.

The F-L .032 calculation is equal to

$$C_1[X_4] + C_2[X_3] + C_3[X_2] + C_4[X_1]$$

which is the capital gains effect on the fourth quarter annual rate of consumption when the *MPS* averaging procedure of equation (A3) is not used

[7] I wish to thank Friend and Lieberman for supplying me with the data and formulas behind their calculation The OTC index, rather than the Standard and Poor's index, can be used to allocate the average capital gains in each quarter of 1963, as F-L have done The implied *MPS* capital gains coefficient still is calculated to be .013

REFERENCES

J. J. Arena, "Capital Gains and the 'Life-Cycle' Hypothesis of Saving," *Amer. Econ. Rev.*, Mar. 1964, *54*, 107–11.

K. G. Bhatia, "Capital Gains and the Aggregate Consumption Function," *Amer Econ. Rev.*, Dec. 1972, *62*, 866–79.

I. Friend and C. Lieberman, "Short-Run Asset Effects on Household Saving and Consumption: The Cross-Section Evidence," *Amer. Econ. Rev.*, Sept. 1975, *65*, 624–33.

F. Modigliani, "Monetary Policy and Consumption," in *Consumer Spending and Monetary Policy: The Linkages*, Fed. Reserve Bank Boston, *Monetary Conference Series*, no. 5, Boston, June 1971, 9–85.

MIT-PENN-SSRC Econometric Model of the United States, unpublished mimeo, Jan. 1973, Wharton Econometric Forecasting Associates.

National Quotation Bureau, OTC Industrial Stock Average, 1962–63.

Firm Output and Changes in Uncertainty

By DONALD V. COES*

Two of the central contributions to the theory of economic behavior under uncertainty are the recent articles in this *Review* by Agnar Sandmo analyzing the competitive firm facing an uncertain output price and by Hayne Leland extending many of Sandmo's conclusions to the monopolistic case.[1] One of their principal results is the demonstration that the optimal output for a risk-averse firm is less facing an uncertain output price than it would be if the firm faced a certain price of the same expected value.

If we recognize that the firm almost always faces some degree of demand uncertainty, then its response to a change in the level of uncertainty is clearly of more empirical interest than is a simple comparison of the certainty and uncertainty case. Sandmo advances the intuitively appealing proposition that a marginal increase in uncertainty will decrease output, but concludes that this conjecture cannot be proved categorically In an interesting recent examination of the risk-averse firm's input demand under output price uncertainty, Raveendra Batra and Armen Ullah have indirectly proved a more restricted version of the Sandmo conjecture for the competitive case, using a production function approach.[2] This note presents a proof of Sandmo's conjecture, using his more general cost function approach and a technique analogous to that of Batra and Ullah, under the widely accepted assumption that absolute risk aversion is nonincreasing. The conclusion is extended to the quantity setting monopolist[3] using the "stochastic demand curve" concept developed by Leland. In the interests of brevity, the principal results of Sandmo and Leland are stated without proof.

It will be assumed that the firm maximizes the expected value of the utility of profits. Its utility function is assumed to be concave (risk averse) and to exhibit nonincreasing absolute risk aversion in the Pratt-Arrow sense $(d[-U''(\pi)/U'(\pi)]/d\pi \leq 0)$.[4] The firm's demand curve is assumed to be an implicit function of price, quantity, and a random element u, which in the quantity setting case may be written as

$$(1)\quad p = p(q,u) \quad (\partial p/\partial q < 0,\ \partial p/\partial u > 0)$$

We assume that the firm has a subjective probability distribution for u, so that by setting q, a conditional distribution of p is then uniquely determined. Following Leland, we assume that as total expected revenue increases, the "riskiness" or dispersion of total revenue increases.

The firm's problem is then to set q so as to maximize the expected utility of profits

$$(2)\quad \max_q E\{U[p(q,u)q - c(q) - b]\}$$

First- and second-order conditions require that

$$(3)\quad E\{U'(\pi)[MR(q,u) - MC(q)]\} = 0$$

$$(4)\quad D = E\{U''(\pi)[MR(q,u) - MC(q)]^2 + U'(\pi)[\partial MR/\partial q - \partial MC/\partial q]\} < 0$$

*Assistant professor, University of Illinois I am grateful to William Branson, Richard Kihlstrom, and an anonymous referee for helpful comments; and to Polly Allen for comments on a earlier version

[1]Leland's notation is followed in this note, permitting Sandmo's model to be expressed as a special case, with $\partial p/\partial q = 0$.

[2]Batra and Ullah assume that the firm's output is produced using a "well-behaved" ($f_{ii} < 0, f_{ij} > 0, i, j = 1,2$) two-factor production function No restrictions are placed on technology, other than the assumption that marginal cost is nondecreasing

[3]If the monopolist sets price rather than quantity, then as Leland shows, optimal price under demand (quantity) uncertainty may be either greater or less than it would be under certainty Hence, without further restrictions, the results presented here do not have a price-setting analogue.

[4]The commonly accepted hypothesis that absolute risk aversion is nonincreasing may be interpreted to mean that as the decision maker becomes wealthier, the risk premium, or the amount he would pay to secure with certainty the expected value of an uncertain prospect, would at least not increase

Leland defines $p = f(q)$, the certainty demand curve equivalent to the stochastic curve $p = p(q,u)$, as the curve the firm would face if it knew with certainty that p would equal its expected value for all q. This implies that expected marginal revenue will equal marginal revenue derived from the certainty demand curve for all q. Following Sandmo and Leland, we may then show that risk aversion requires that under uncertainty

$$(5) \qquad E\{U'(\pi)[MR(q,u) - MR(q,u_a^1)]\} \leq 0$$

where u_a^1 is the u at which marginal revenue equals its expectation or certainty equivalent for a given q. Inequality (5) states that the expected value of marginal utility-weighted deviations of actual marginal revenue from their expectation is nonpositive, a consequence of the fact that positive but decreasing marginal utility assigns relatively greater weight to negative deviations from expected marginal revenue than to positive ones.

Rewriting the first-order condition (3) as $E\{U'(\pi)MR(q,u)\} = E\{U'(\pi)MC(q)\}$ and subtracting $E\{U'(\pi)MR(q,u_a^1)\}$ from both sides, we have from (5)

$$(6) \quad MR(q,u_a^1) = E\{MR(q,u)\} \geq MC(q) \quad \text{for all } q$$

Inequality (6) states that optimal output under uncertainty will be chosen so that marginal cost is less than or equal to expected marginal revenue. If marginal costs are nondecreasing in q, this implies that optimal output will be less than or equal to its level under certainty.

Noting that $MR(q,u_a^1) = f(q) + q(\partial f(q)/\partial q)$, we may rewrite (5) as

$$E\{U'(\pi)[MR(q,u) - f(q)]\} - E\left\{U'(\pi)\, q\, \frac{\partial f(q)}{\partial q}\right\} \leq 0$$

The second term is nonpositive, since $U'(\pi) > 0$, so that

$$(7) \qquad E\{U'(\pi)[MR(q,u) - f(q)]\} \leq 0$$

Under the assumption of nonincreasing absolute risk aversion, Sandmo shows that

$$(8) \quad E\{U''(\pi)[MR(q,u) - MC(q)]\} \geq 0$$

This inequality, unlike (5), does not have an intuitive economic interpretation. Some insight, however, is provided by the fact that by its definition, nonincreasing absolute risk aversion is sufficient for $U'''(\pi) > 0$. For small u, the term $MR(q,u) - MC(q)$ will be negative and weighted relatively heavily by $U''(\pi)$, while for large u, this term is positive but $U''(\pi)$ tends toward zero.

A pure change in the level of uncertainty may be defined as a change in u such that the conditional distribution of p is spread about its certainty equivalent for any q, leaving its expectation unchanged.[5] Letting the price after the increase in uncertainty be $p^*(q,u^*) = \gamma p + \theta$, where γ and θ are multiplicative and additive parameters initially equal to unity and zero, respectively, we then require

$$dE[p^*(q,u^*)] = E[p(q,u)]d\gamma + d\theta = 0$$

or

$$(9) \qquad \frac{d\theta}{d\gamma} = -E[p(q,u)] = -f(q)$$

The first-order condition then becomes

$$(3') \quad A = E\{U'[(\gamma p + \theta)q - c(q) - b] \cdot [(\gamma p + \theta) + \gamma \frac{\partial p}{\partial q} q - c'(q)]\} = 0$$

which expresses q as an implicit function of the multiplicative shift parameter γ. Differentiating with respect to γ, setting $\gamma = 1$ and $\theta = 0$, and using (4) and (9), we have

$$(10) \quad \frac{dq}{d\gamma} = \frac{-1}{D} E\{U'(\pi)[MR(q,u) - f(q)]\} + \frac{-1}{D} E\{U''(\pi)[q(p - f(q)) (MR(q,u) - MC(q)]\}$$

By (4) and (7) the first term is nonpositive, so

[5]The multiplicative spreading of the distribution of the random variable as a definition of an increase in uncertainty was introduced by Sandmo and is a particular case of a more general class of "mean-preserving spreads" investigated by Michael Rothschild and Joseph Stiglitz.

that a sufficient condition for $dq/d\gamma < 0$ is that the second term be negative. As $MR(q,u)$ is monotonic in p, we have

$$(11) \qquad MR(q,u) \gtreqless MR(q,u_a^1) \quad \text{iff} \quad p(q,u) \gtreqless p(q,u_a^1) = f(q)$$

Hence the second term of (10) will have the same sign as the expression

$$\frac{-1}{D} q\, E\{U''(\pi)[MR(q,u) - MR(q,u_a^1)][MR(q,u) - MC(q)]\}$$

which may be expanded to give

$$\frac{-1}{D} q\, E\{U''(\pi)[MR(q,u) - MC(q)] \cdot [MR(q,u) - MC(q) + MC(q) - MR(q,u_a^1)]\}$$

or

$$\frac{-1}{D} q\, E\{U''(\pi)[MR(q,u) - MC(q)]^2\} + \frac{-1}{D} q\, E\{U''(\pi)[MR(q,u) - MC(q)] [MC(q) - MR(q,u_a^1)]\}$$

The first term in this expression is unambiguously negative, since $U''(\pi) < 0$ for all q,u. By (6), expected marginal revenue at any q is greater than or equal to marginal cost, so that the nonstochastic term $MC(q) - MR(q,u_a^1)$ is nonpositive for all q. With (7), this requires that

$$(12) \qquad \frac{dq}{d\gamma} < 0 \qquad \text{for all } q$$

An increase in demand uncertainty, which may be defined as a "stretching" of the conditional distribution of p about a constant certainty equivalent $f(q)$ decreases the optimum output of the quantity-setting firm, whether it is a perfect competitor or a monopolist. A sufficient, although not necessary condition for this to hold is that absolute risk aversion be nonincreasing.[6]

[6]The assumption of decreasing absolute risk aversion used by Batra and Ullah is clearly an overly strong sufficient condition for $dq/d\gamma < 0$ If $R_a'(\pi) = 0$, (8) becomes an equality, eliminating only one of three nonpositive terms in the expanded form of (10).

REFERENCES

R. Batra and A. Ullah, "Competitive Firm and the Theory of Input Demand under Price Uncertainty," *J. Polit. Econ.*, May/June 1974, *82*, 537–48.

H. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

J. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 127–36.

M. Rothschild and J. Stiglitz, "Increasing Risk: I. A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43.

A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

Academic Achievement and Job Performance: Note

By EDWARD LAZEAR*

David Wise's recent paper in this *Review* addresses two questions: First, is there any relationship between the subjective quality of an individual's college institution or his relative position within that college and his eventual job performance? Second, if a relationship exists, what is its causal nature? Specifically, do variations in schooling types affect productivity directly, or are they merely associated with higher productivity through screening channels? Although Wise's analysis yields a convincingly affirmative answer to the first question, it is silent on the second. The "indirect evidence" that Wise offers is quite consistent with the screening as well as the productivity-augmenting hypothesis. His conclusion that college education contributes to productive ability is therefore unwarranted.

The screening hypothesis in its most basic form[1] asserts that schooling acquisition costs differ across individuals according to their ability levels. If high ability individuals face lower marginal cost of schooling schedules than do low ability individuals, the former group will for a given return obtain more education. Employers will pay higher wages to the more educated because they recognize that ability and attained level of education are positively correlated as the result of differential costs.

Screening can be contrasted with the productivity augmentation view of schooling. The latter position holds that schools actually do alter an individual's productivity not simply by producing an optimal sort (although this is not excluded),[2] but primarily by augmenting an individual's *ex post* ability.

What is important here is a fact that has made it virtually impossible to come up with a valid test of the screening hypothesis. That fact is that from an individual's point of view, it is almost always irrelevant whether schooling is a screen or productivity augmenter.[3] Human capital analyses are consistent with both. Since the individual is simply assumed to maximize the present value of his income stream, he is unconcerned with the employer's reason for paying higher wages. As long as the acquisition of schooling is the least expensive way to inform potential employers of his high ability, he will still "invest" in it just as he would if it actually increased his productivity (In fact, the two hypotheses have different implications for society only when there are lower cost ways to the group of providing information on differential ability. The possibility arises in Spence when individuals know their true abilities and can be more cheaply induced to report them accurately). This brings us back to the Wise paper which, like others,[4] attempts to ascertain the validity of the screening hypothesis

Wise cites four pieces of evidence which he claims lend support to the productivity-augmenting view of schooling. First, he argues that if grades merely attest to a student's *ex ante* ability rather than to differences in amounts of acquired knowledge (under screening, this is zero for all individuals), grades should have a small effect on compensation for individuals from higher quality colleges. (Students from these schools, he suggests, vary less in innate ability than those from lower quality institutions). The "screening prediction" is not borne out. Nor should it be. Even if differences in ability are finer on the scale of academic per-

*Assistant professor of economics, University of Chicago, and research associate of the National Bureau of Economic Research.

[1]See Michael Spence or Joseph Stiglitz for a clear discussion of the screening hypothesis

[2]See Finis Welch (1970, 1973) for a variant on this theme.

[3]A possible exception is mentioned below.

[4]See Kenneth Wolpin, for example. This is discussed briefly below

formance (such as SAT scores), there is no reason to expect that the mapping from school performance or even "innate" ability to worker productivity and compensation is linear.[5] That is, differences in "ability" at high levels may matter more for work-productivity variations than differences in ability at low levels where jobs command less responsibility. Differences in grades at high levels of ability, i.e., where individuals graduate from highly rated institutions, may well be associated with larger wage level and wage growth differences than those at low levels of ability. This result would be quite consistent with screening.

Wise's second argument relates to incremental effects of graduate education. If grades are merely a stamp which certify the individual's underlying attributes, he suggests, there should be no additional impact of graduate school performance on salary once undergraduate records are held constant. His evidence contradicts this prediction. But this can hardly be construed to be inconsistent with screening. All one has to argue is that both pieces of information are useful to an employer's evaluation. Nor does this require that one's imagination be stretched. There is no reason why an individual who obtains an A average as an undergraduate, but a C average in his M.A. program should have the same expected "innate" ability as the student who received A's throughout his education. The fact that an individual is willing to undertake additional schooling may, by itself, say something about inherent skills. His performance in that graduate program merely tells more.

A related piece of evidence deals with high school versus college performance. The prediction of a screening-type explanation, Wise claims, is that the effects of high school grades and undergraduate records should be confounded. Instead, he finds that high school grades are insignificant and the effect of college grades is virtually unaffected. This does not refute the screening hypothesis. Rather, it says that high school performance as measured by grades does not matter. This finding, if disturbing, is disturbing to the productivity-augmenting hypothesis as well. The insignificance of high school grades probably results from another factor. College quality might say much more about high school performance through selectivity than high school grades themselves. If college quality is included in the (unreported) regression, the insignificance of high school grades would be less than surprising.

The final point relates to the finding that there is little correlation between college quality or rank within the college and initial position within the firm. If we take the evidence at face value, it should suffice to point out that this poses a problem for productivity-augmenting interpretations of schooling as well. This puzzle, along with the corresponding finding that initial salary is unaffected by these factors, can be reconciled. If schooling is merely a screen, then upon entry to the firm, individuals may differ, not in their current productivities, but in their abilities to acquire job-related skills at work. Thus, initial wages would be similar while rates of wage growth would differ

The point may be generalized. It is often suggested that if schooling acts merely as a screen, it should become relatively less important as one's working life progresses. Thus, differences in wages between schooling groups should be most important upon graduation. However, inferences on screening drawn from this sort of analysis are incorrect. First, even if schooling is not productive per se, the more able, highly schooled individuals may enjoy steeper age-earnings profiles simply because of differential ability growth or depreciation over time. For example, age may have a greater detrimental effect on the productivity of low ability workers who specialize in physical activity than on high ability workers who specialize in mental activity. Second, if innate ability affects the rate of skill acquisition once on the job, more highly schooled workers can anticipate more rapid wage growth as the result of higher returns to

[5]Furthermore, there is little reason to even accept the assumption that abilities vary less at high quality schools As Wise points out, SAT scores are at best ordinal If ability is distributed normally, the differences between the highest and lowest score in the top and bottom deciles will be larger than the difference between the highest and lowest score in a middle decile

on-the-job training. This is consistent with the notion of job tracking. If individuals get on job tracks, their perceived initial ability for which schooling may be a signal, can be instrumental in altering the slope of age-earnings relationships.[8]

Additional rationalizations are numerous, but the point is clear. Inferences cannot easily be drawn from wage growth data to either refute or confirm screening explanations of schooling. As already mentioned, the criticism should not be confined to the Wise paper. Attempts to test for the validity of screening have been frustrated elsewhere as well. Therefore, consider two potential tests of screening.

The first is derived from the observation that employers sometimes offer to send employees to school "all expenses paid." If screening were the only motivation for schooling acquisition, one would not observe employers bearing all costs since the screening mechanism works by forcing individuals to bear differential costs of schooling. However, what one observes is that workers in a given job at a given wage receive the same nominal compensation to attend school. Even in the context of screening, those with the highest ability will have the lowest true cost and therefore earn the highest profit on this transaction. Thus, the more able are more likely to accept the employer's schooling offer and the screening hypothesis is saved again.

The second test relies on finding a group of individuals for whom screening is irrelevant. Wolpin uses this approach. He argues that the self-employed will not invest in schooling if it is merely a screen. Thus, under screening, self-employed individuals should have lower attained levels of education. He finds that they do not. This test requires two assumptions: First, it must be the case that individuals determine the probability of self-employment before investing in education. Second, it must be the case that their customers do not use their credentials as a signal in assessing product quality. In the case of physicians, dentists, lawyers, and other professionals, the second assumption is unlikely to be valid.

Valid methods for separating screening from productivity-augmenting views of education on the basis of different implications are hard to come by. The Wise paper, although perhaps misnamed, is interesting and important in providing evidence on differential returns to schooling. It cannot, however, be regarded as a refutation of the screening hypothesis.

[8]This is consistent with queue theories of labor markets. Here the most able are most likely to be trained. See (among others) Lester Thurow for a discussion along these lines.

REFERENCES

M. Spence, "Job Market Signalling," *Quart. J. Econ.*, Aug. 1973, *87*, 355–79.

J. Stiglitz, "The Theory of Screening, Education and the Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

L. Thurow, "Education and Economic Equality," *Publ. Interest*, Summer 1972, *28*, 66–81.

F. Welch, "Education in Production," *J. Polit Econ.*, Jan./Feb. 1970, *78*, 35–59.

———, "Education, Information, and Efficiency," Nat. Bur. Econ. Res. working pap. no. 1, June 1973.

D. Wise, "Academic Achievement and Job Performance," *Amer. Econ. Rev.*, June 1975, *65*, 350–66.

K. Wolpin, "Education and Screening," Nat. Bur. Econ. Res. working pap. no. 104, Aug. 1975.

ERRATA

On the Shape of the Trade Indifference Curve: Rejoinder to Batra

By MURRAY C. KEMP AND EDWARD TOWER*

Please note the following revision in the September 1976 issue of this ***Review:***

Page 709:

The first sentence of the second paragraph should be corrected to read:

"Let $C = \{c = (c_1, c_2, \ldots, c_n) \in R^n_+ \mid U(c) \geq U_0\}$ be the set of all consumption bundles which generate levels of utility not less than U_0, and let $Y = \{y = y_1, y_2, \ldots, y_n) \in R^n_+ \mid (y, \bar{x}) \in T\}$ be the production possibility set, where T denotes the set of input-output combinations that are feasible and $\bar{x}$ denotes the given factor endowments.

*Research professor of economics, University of New South Wales, and associate professor of economics, Duke University, respectively

NOTES

The Association of Indian Economic Studies will hold its next conference, "New Directions in India's Economic Development," Aug. 19–21, 1977 at Montclair State College. Please send proposals for papers and participation to Professor Suresh Desai, Economics Department, Montclair State College, Upper Montclair, NJ 07043.

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies Grants are likely to cover only lowest cost excursion fares and will rarely exceed 50 percent of full economy-class fares. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical Conferences dealing with the establishment of social policy or legislation are ineligible The deadlines for applications to be received in the office of the American Economic Association are. meetings scheduled between July and October, March 1, for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1 Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212

The National Tax Association-Tax Institute of America announces the 1976 award winners in the annual competition for outstanding doctoral dissertations in government finance and taxation The $1,000 first prize award was won by Patrick D. Larkey of the University of Michigan (now at the University of British Columbia) with his entry, "Process Models and Program Evaluation The Impact of General Revenue Sharing on Municipal Fiscal Behavior " Honorable mention awards of $500 each were won by Thomas S McCaleb of the University of North Carolina, (now at the University of Kansas), "Optimal Income Taxation An Integration of Private and Public Choice Considerations" and Nonna Anne Noto of Stanford University (now at the Federal Reserve Bank of Philadelphia), "The Influence of the Local Public Sector on Residential Property Values " The members of the 1976 Selection Committee were Professors Harvey E Brazer, Arthur D Lynn, Oliver Oldman, and James A Papke. Information on the 1977 award competition may be obtained from Professor James A Papke, Department of Economics, Krannert Graduate School of Management, Purdue University, West Lafayette, Indiana 47907

For thirty years the Fulbright-Hays program has provided opportunities for university lecturing and advanced research abroad In recent years 450–500 awards per year have been made to American scholars and other professionals—25–30 to specialists in economics and business. The program also includes awards to foreign scholars for lecturing and advanced research at *U.S.* institutions Announcement of the awards available for 1978–79, in the 31st annual competition, will be published in March 1977. The general composition of the program involving more than 70 countries is expected to be similar to that of recent years. Registration for personal copies of the announcement is now open, forms are available from the Council for International Exchange of Scholars, Suite 300, Eleven Dupont Circle, Washington, D.C 20036

An all-day conference on post-Keynesian theory will be held on Saturday, Apr 16, at Rutgers University For further information, contact either Professor Alfred S. Eichner, State University of New York, Purchase, NY 10577, or Professor Paul Davidson, Department of Economics, Rutgers University, New Brunswick, NJ 08903

The Institute of Public Policy Studies at the University of Michigan has received support from the National Institute of Mental Health for postdoctoral training in public policy analysis Training will take the form of course work and/or research in policy analysis Candidates interested in social problems related to mental health (urban problems, minority group problems, delivery of public services, etc) or in evaluation research are urged to apply for stipends ranging from $10,000 to $13,200 Applications should include a statement on proposed research program, two or more letters of recommendation, and a transcript of work completed at the doctoral level Send applications or inquiries to Professor Joel D Aberbach, Institute of Public Policy Studies, 318 Gunn Bldg., 506 E. Liberty Street, Ann Arbor, MI 48109 (313+763-4212).

The National Science Foundation's Office for the International Decade of Ocean Exploration (*IDOE*) has established a Marine Science Affairs program to support research on the social, economic, political, and managerial implications of the *IDOE* scientific program This program consists of long-term, multidisciplinary oceanographic projects in four major programs, environmental quality, environmental forecasting, seabed assessment, and living resources The program will begin with about $200,000 in fiscal year 1977 Additional information may be obtained by contacting Program Manager, Marine Science Affairs Program, Office for the International Decade of Ocean Exploration, National Science Foundation, Washington, D.C. 20550 (202+632-7356).

Most projections of supply and demand for Ph.D.s in science and engineering estimate that the number of new doctoral recipients will provide about one-third more Ph D s by 1985 than required for academic and research openings. To address some of the problems inherent in this projected surplus, the Higher Education Research Institute (*HERI*) has launched a project, funded by the Ford and National Science Foundations, to identify nontraditional job markets for academics and determine entry points for new Ph D. recipients *HERI* will survey those in nontraditional jobs to determine why doctorate holders take such employment and to evaluate job satisfaction and other career outcomes The study will compare backgrounds, training, and attitudes of those in traditional and nontraditional careers. The study will sample 15,000 doctorate holders from 14 science (physical, biological, and social) and engineering fields who have moved out of or into nonacademic jobs in the last 3 years. *HERI* will also survey Ph.D.s who have moved between academic institutions, from academic or research to administrative positions within a college, university, or research organization, or within or between business firms The study will explore the career patterns, characteristics, and routes of access of the highly mobile Ph D s

If you hold a Ph.D in economics and hold a nontraditional or unusual job outside the academic or traditional research areas, or have changed employers or job functions within the past 3 years, and would like to participate in the study, send your name, address, field of study, and year you completed your Ph.D to Higher Education Research Institute, 924 Westwood Blvd, Suite 850, Los Angeles, CA 90024.

BOOKS FOR ASIA, a project of The Asia Foundation, asks that you send books and journals you are no longer using to the address given below. Books must be published in 1965 or later, and be in excellent condition. At least one complete year of a journal published since 1950, and long complete runs in particular, are needed Donations of books and journals are tax deductible If you have any questions or wish to send materials, please direct them to Books for Asia, Attn Carlton Lowenberg, Director, 451 Sixth Street, San Francisco, CA 94103 (415+982-4640)

Members of the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics are pleased to announce the institution of an annual Leonard J. Savage Award of five hundred dollars ($500) for an outstanding doctoral dissertation in the area of Bayesian Econometrics and Statistics To be considered for the 1977 Savage Award, a doctoral dissertation must be submitted by the dissertation supervisor before July 1, 1977 and accompanied by a short letter from the supervisor summarizing the main results of the dissertation. Dissertations completed after Jan 1, 1976 are eligible to be considered for the 1977 Savage Award. An Evaluation Committee will be appointed by the board of the Leonard J Savage Memorial Trust Fund (S. E Fienberg, S Geisser, J B. Kadane, E. E Leamer, J. W Pratt, and A Zellner, chairman) to evaluate dissertations that are submitted for the Savage Award Dissertations and supporting letters should be sent to: Professor Arnold Zellner, Graduate School of Business, University of Chicago, 5836 S. Greenwood Avenue, Chicago, Illinois 60637.

House Subcommittee Seeks Views on Emergency Trade Legislation

The House Subcommittee on International Trade and Commerce will conduct studies leading to public hearings during the first six months of 1977 on the advisability of revising or repealing section 5(b) of the Trading With the Enemy Act of 1917. This section authorizes the President to exercise extensive controls over international economic transactions in time of declared national emergency. Since the nation has been in a declared state of emergency since 1933, these authorities are in effect available for the day-to-day conduct of foreign policy They are currently used by the administration to impose trade embargoes on North Korea, Vietnam, Cambodia, and Cuba, and for routine export controls normally exercised under the temporarily lapsed Export Administration Act.

The Subcommittee recently published a complete legislative and administrative history of section 5(b) entitled, "Trading With the Enemy. Legislative and Executive Documents Concerning Regulation of International Transactions in Time of Declared National Emergency " This volume is available in limited quantities from the Committee on International Relations, House of Representatives, Washington, DC. 20515, and for $2 75 from the Government Printing Office, Washington, D.C 20402

The Subcommittee would like to hear from scholars and other interested people engaged in analyzing possible changes in U S. laws to provide a better basis for both routine and emergency foreign economic and trade policy People doing work or wishing to present views on U.S. foreign economic policy, particularly with reference to emergency trade controls, asset controls, trade embargoes, and related legal and policy questions, are invited to communicate with the Subcommittee at 707 H.O B Annex No. 1, House of Representatives, Washington, D.C 20515 (202+725-3246)

Call for Abstracts

The Conference on Social Sciences in Health is seeking a limited number of contributed papers for presentation as part of its program at the meetings of the American Public Health Association in Washington, D C on Oct 30–Nov. 3, 1977 Preference will be given to policy relevant papers which reflect a significant social science perspective. Interested persons should submit an abstract of up to 200 words by Apr. 15, 1977 Persons whose abstracts are selected for further consideration will be asked to submit a three page summary Please send abstracts to Dr. Irving Leveson, Senior Professional Staff, Hudson Institute, Quaker Ridge Road, Croton-on-Hudson, N.Y 10520.

Deaths

Carl A. Dauten, Executive Vice Chancellor, Washington University, St. Louis, Sept. 17, 1976

M. H. Dobb, Aug. 19, 1976.

Walter G. Miller, retired economist, Federal Highway Administration, Department of Transportation, Oct 18, 1976.

E Bryant Phillips, professor emeritus, department of economics, University of Southern California, Dec. 19, 1975.

Vladimir Stoikov, New York State School of Industrial and Labor Relations, Cornell University, Aug 2, 1976.

Philip Taft, professor emeritus, Brown University, November 17, 1976.

Retirements

Albert Abrahamson, professor of economics, Bowdoin College, June 1976.

Richard V. Clemence, professor of economics, Wellesley College, June 1976

Daniel C Kaufherr, American Graduate School of International Management, June 30, 1976.

Harald Leuba, American Graduate School of International Management, May 1, 1976

Alvin M. Marks, American Graduate School of International Management, June 30, 1976.

George Peck, professor, department of economics, University of Iowa, Aug 1976.

Roderick H. Riley, consulting economist (since June 1, 1971, retired economic advisor, Bureau of Indian Affairs, U.S. Department of the Interior) member, Maryland Property Tax Assessment Appeal Board for Montgomery County, Dec 18, 1975.

Visiting Foreign Scholars

Urie Ben-Zion, Israel Institute of Technology visiting associate professor of economics, University of Iowa, Jan. 1976.

Walter Eltis, Exeter College, Oxford: visiting professor, department of political economy, University of Toronto, until Sept. 1977

Erich Klinkmuller, Freie Universität Berlin. visiting professor, international studies, American Graduate School of International Management, fall 1976

Hukukane Nikaido, Hitotsubashi University, Tokyo department of economics, University of Southern California, spring 1977.

Promotions

Guenter M. Conradus. vice president, economics, Mathematical Sciences Northwest, Inc., Sept. 1976

Stuart M. Feder: chief, International Reports Division, Federal Reserve Bank of New York, Sept. 2, 1976.

Joseph E. Haring: professor of economics. Occidental College, July 1, 1976.

Calvin A. Hoerneman: professor of economics, Delta College.

Joseph M. Jadlow. professor, department of economics, Oklahoma State University.

Pauline W Kopecky. associate professor, department of economics, Oklahoma State University.

Gerald M. Lage: professor, department of economics, Oklahoma State University

Peter R. Moore: associate professor, department of economics and management, Rhode Island College, July 1976

Rodney J. Morrison: professor of economics, Wellesley College, Sept. 1976.

Jeffrey B. Nugent. professor of economics, University of Southern California, Sept. 1, 1976.

John D. Rea. associate professor, department of economics, Oklahoma State University.

James D. Rodgers. professor of economics, Pennsylvania State University, July 1, 1976

John M. Sapinsley. associate professor, department of economics and management, Rhode Island College, July 1975

Robert C Stuart. professor of economics, Douglass College, Rutgers–The State University, July 1, 1976

Arnold H. Studenmund: associate professor, Occidental College, July 1. 1976.

David J Vail: associate professor of economics, Bowdoin College, Sept 1976

Administrative Appointments

Robert P Collier, Utah State University· dean of the College of Business and Economics, Western Washington State College

Paul G Darling. chairman, department of economics, Bowdoin College, Sept 1976

Hugh N Emerson, University of Maryland. chairman, business and economics department. Alderson-Broaddus College, Sept 1, 1976.

Richard H. Day, University of Wisconsin-Madison chairman, department of economics, University of Southern California, Sept. 1, 1976.

Robert J McMahon. chairman, department of world business, American Graduate School of International Management, June 1, 1976.

Peter R Moore. chairman, department of economics and management, Rhode Island College, July 1976

Tapan Munroe: chairman, department of economics, University of the Pacific, Feb. 1, 1976

Harold J Noah dean, Teachers College, Columbia University, Sept. 1, 1976.

Robert C Stuart· chairman, department of economics, Douglass College, Rutgers–The State University, July 1, 1976.

Richard O Zerbe: acting director, Program in Social Management of Technology, University of Washington, Sept. 16, 1976.

Appointments

Akgar Ahktar: economist, Foreign Research Division, Federal Reserve Bank of New York, Aug. 31, 1976.

Dennis J. Aigner, University of Wisconsin-Madison. professor, department of economics, University of Southern California, Sept 1, 1976.

Narongchai Akrasanee: visiting research fellow, National Bureau of Economic Research, Aug 1976.

E. Jane Arnault: assistant professor, Occidental College, July 1, 1976.

Thomas D. Boston: assistant professor, department of economics, Atlanta University and Clark College, Sept 1, 1976.

Thomas H Bruggink, University of Illinois, instructor, department of economics, Fordham University, Sept. 1, 1976.

Karl Case, Harvard University, instructor in economics, Wellesley College, Sept. 1976.

Alexander H. Cornell: associate professor, department of economics and management, Rhode Island College, Sept 1975.

W. Davis Dechert, Cornell University: instructor, department of economics, University of Southern California, Sept 1, 1976.

Ronald E. Deiter: assistant professor, department of economics, Iowa State University, Sept. 1, 1976.

C Frederick DeKay, senior economist, Mathematical Sciences Northwest, Inc., Mar. 1976.

Nancy J DelPrete instructor, department of economics and management, Rhode Island College, Sept 1976

Jonathan Dickinson, assistant professor of economics, Pennsylvania State University, Sept. 1, 1976

David W. Dunlop, Vanderbilt University: visiting assistant professor, department of economics, Dartmouth College, fall 1976

Richard F Dye, assistant professor, department of economics, Bowdoin College, fall 1976

Robert M Feinberg: assistant professor of economics, Pennsylvania State University, Sept 1, 1976

John Fox: assistant professor of economics, University of South Carolina, Sept 1976.

John C Goodman, Columbia University visiting assistant professor, department of economics, Dartmouth College, 1976–77

Eileen C Gram, Barnard College assistant professor, department of economics, Fordham University, Sept 1, 1976

Daniel C Green assistant professor of statistics, American Graduate School of International Management, Sept 1, 1976

Manuel d J Gutierrez, Webster College, instructor of applied statistics, Stockton State College, Pomona.

R Duane Hall associate professor of marketing, American Graduate School of International Management, Sept 1, 1976

Bryan Heathcott: assistant professor of finance, American Graduate School of International Management, Sept. 1, 1976

C Michael Henry: instructor, department of economics, Rutgers–The State University, July 1, 1976

Edward A. Holt: senior resources planner, Mathematical Sciences Northwest, Inc., July 1976

Stephen M Horner, University of Michigan: instructor in economics, Wellesley College, Sept. 1976

Herbert L. Johnson: assistant professor, department of economics and management, Rhode Island College, Sept 1976.

Nake M. Kamrany, adjunct professor, department of economics, University of Southern California, Sept. 1, 1976

Richard J. Kent, University of California, Berkeley: assistant professor, department of economics, Dartmouth College, Sept. 1976.

Robert D. Lamson: principal economist, Mathematical Sciences Northwest, Inc., Nov. 1976.

Stephen J. Land: senior economist, Mathematical Sciences Northwest, Inc., May 1976.

William M. McHugh, senior economist, Mathematical Sciences Northwest, Inc., August 1976.

Nancy P. Marion, Princeton University: instructor, department of economics, Dartmouth College, Sept. 1976.

Philip Mayer, instructor, department of economics, Southwest Missouri State University, Aug. 16, 1976.

William C Melton: economist, Money and Finance Division, Federal Reserve Bank of New York, Aug. 3, 1976.

Emad Mohit: instructor of economics, American Graduate School of International Management, Sept. 1, 1976.

James F. O'Connor, assistant professor, department of economics, University of Iowa, Aug. 1976

Stephen Owuso: economist, Business Conditions Division, Federal Reserve Bank of New York, Sept. 2, 1976

Nils A Parr: associate professor of economics, Environmental Studies Program, University of Maine at Machias, Sept 1976

Harold Payson III: assistant professor, department of economics, Bowdoin College, fall 1976.

Joseph Pelzman: assistant professor of economics, University of South Carolina, Sept. 1976.

Joel Popkin: director, National Bureau of Economic Research, Washington

John Rapoport, Mount Holyoke College: visiting assistant professor, department of economics, Dartmouth College, 1976–77

Jonathan Ratner, Yale University: instructor in economics, Wellesley College, Sept. 1976

Raymond Riezman, instructor, department of economics, University of Iowa, Aug 1976.

Joanna Robinson, University of Connecticut: assistant professor of economics, Wellesley College, Sept. 1976

Paul J Schlesinger: associate professor of advertising and marketing, American Graduate School of International Management, Feb 1, 1976.

Marjorie H Schnader economist, Business Conditions Division, Federal Reserve Bank of New York, Sept 14, 1976

Stuart O Schweitzer, Georgetown University: associate professor of public health, University of California, Los Angeles, Sept 1976.

Eugene Short, economist, Foreign Research Division, Federal Reserve Bank of New York, Oct. 5, 1976.

Joseph F. Sinkey, Jr. associate professor, department of banking and finance, College of Business Administration, University of Georgia, Sept 1, 1976.

Sharon Smith, economist, Business Conditions Division, Federal Reserve Bank of New York, Sept. 7, 1976.

Stanley P. Stephenson, Jr. assistant professor of economics, Pennsylvania State University, Sept. 1, 1976.

Kenneth E. Stone, assistant professor, department of economics, Iowa State University, Sept. 1, 1976.

Richard C. Tepel, Brown University: assistant professor, department of economics, Fordham University, Sept. 1, 1976.

Thom B. Thurston: economist, Market Statistics Division,

Federal Reserve Bank of New York, Sept. 2, 1976.

George M. Wattles: professor of world business, American Graduate School of International Management, Sept. 1, 1976.

D. Sykes Wilford: economist, Foreign Research Division, Federal Reserve Bank of New York, Aug. 24, 1976.

Deloris R. Wright: assistant professor, department of economics, Southwest Missouri State University, June 7, 1976.

Edward Yardeni: economist, Money and Finance division, Federal Reserve Bank of New York, Sept 16, 1976.

James N. Young, Georgia State University: assistant professor, department of economics, Fordham University, Sept. 1, 1976.

Jeffrey T. Young, University of Colorado-Denver. assistant professor of economics, Marshall University, Aug 30, 1976.

Michael Zubkoff: adjunct professor, department of economics, Dartmouth College

Leaves for Special Appointments

Arthur G. Ashbrook, Jr., U S government, CIA visiting professor of economics, U.S Naval Academy, Sept 1, 1976

Louis J. Cherene, Jr., University of Southern California: University of Hull, Great Britain, 1976–77.

Warren Dent, University of Iowa: visiting associate professor, University of Wisconsin-Madison, 1976–77.

Richard Weisskoff, Iowa State University: research fellowship, Social Science Research Council, Southern Peru, Sept. 1, 1976–Feb. 28, 1977.

Resignations

William J. Barger: University of Southern California, Aug 31, 1976.

Bernard H Booms, Pennsylvania State University: Harvard University, June 30, 1976.

Robert J Latham, Pennsylvania State University: Iowa State Commerce Commission, June 30, 1976

William H. Peterson, American Graduate School of International Management, June 29, 1976

Gordon C Rausser, Iowa State University: Harvard University, Aug. 31, 1976.

Phillip E. Vincent, University of Southern California, Jan 31, 1976.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories.

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B Please give the name of the individual (SMITH, John W), his present place of employment or enrollment, his new title (if any), and the date at which the change will occur.

C Type each item on a separate 3x5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows *March*, November 1, *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.